PROCEEDINGS

OF THE

NATIONAL ACADEMY OF SCIENCES

OF THE

UNITED STATES OF AMERICA

VOLUME 31, 1945

PUBLISHED MONTHLY

Publication Office: Mack Printing Company, Easton, Pa.

Editorial Office: Harvard School of Public Health, Boston 15

Home Office of the Academy: Washington 25, D. C.

CONTENTS

PHYSICS

ENGINEERING

CHEMISTRY

BOTANY

ZOÖLOGY AND ANATOMY

PHYSIOLOGY AND BIOCHEMISTRY

PATHOLOGY AND BACTERIOLOGY

GENETICS

PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 31 | January 15, 1945 | Number 1

THE AMOUNTS OF ACETYLCHOLINE IN THE DARK SKIN AND IN THE PALE SKIN OF THE CATFISH*

By George H. Parker, John H. Welsh and Jane E. Hyde

Biological Laboratories, Harvard University

Communicated November 24, 1944

The dark and the pale phases of the common catfish, *Ameiurus nebulosus*, are brought about by the activity of a single kind of integumentary color-cell, the melanophore, which is innervated by two sets of autonomic nerve-fibres (Parker, 1934). Preliminary evidence has been advanced to show that these two sets of fibres, like those associated with the vertebrate heart, are in one case adrenergic and in the other cholinergic. By the use of organic indicators the pale skin of the catfish has been shown to contain adrenaline and the dark skin acetylcholine (Ach), and by a determination made in the Harvard Laboratories the dark skin of the catfish was found to contain 0.078 gamma of Ach per gram of moist skin (Parker, 1940). At the same time that this determination was being worked out at Harvard three Chinese physiologists, Chang, Hsieh and Lu (1939), published a similar determination made on the skin of the Chinese snake-fish by the same method as was being used on the catfish, a determination which yielded 0.077 gamma of Ach per gram of skin. The very close agreement between the figures for these two determinations must be in a measure accidental. Nevertheless their agreement indicates beyond a doubt the approximate order of magnitude of the amount of Ach involved and shows that this amount is very small.

Ach is now well known to darken many fishes by inducing in their melanophores a dispersion of pigment. Presumably extremely small amounts of this agent are normally released from cholinergic nerves in effecting this dispersion. This would be deduced from the small total amount of Ach extractible from dark fish skin and the usual great difference between the residual Ach of a nervous tissue and the amounts released on its activation (MacIntosh, 1938). However, in spite of this, and with recognition of the efficiency with which Ach reserves may be maintained in active nerves (Brown and Feldberg, 1936), it was believed that extended exposures of fishes to white and to black backgrounds might re-

sult in a measurable difference in the Ach content of the pale and the dark skins. If so, this would yield additional evidence for the view that the melanophore pigment is normally dispersed through the release of Ach. The results of attempts to demonstrate such differences are presented in this paper.

In approaching this problem it might seem at first sight a very simple matter to repeat on the pale skin of a catfish the same kind of test that had been used on the dark skin of this animal. But such is not true, for when the skin of a pale fish is removed preparatory to grinding it, it quickly darkens and thus passes into the opposite color phase from that which it was planned to test. This darkening is due to the stimulation by cutting of the dispersing nerves distributed to the melanophores of the ablated skin. Hence the ordinary removal of pale skin quickly changes it in such a manner as to defeat at the very outset the possibility of the required test. To avoid this difficulty a method must be devised to remove the skin from a pale fish and to keep it in the pale state for the test. To this end several attempts were made in all of which the prime object was to kill the pale skin of the catfish before it had begun to darken and thus to obtain a preparation unquestionably pale in phase.

In the first of these attempts catfishes that had been rendered pale by a sojourn of several days in a white-walled, illuminated tank were taken singly from the tank, decapitated at once with heavy shears, and plunged for fifteen seconds in water at 60°C. By this procedure the skin with its contained melanophores and nerves was quickly killed. It could then be stripped from the fish's body, pulped in a mortar and extracted for testing against standard solutions of Ach with the clam's heart as an indicator; a method based on the findings of Prosser (1940). The critical periods of time from the instant the fishes were lifted from the tank till their skins were killed by the hot water varied from three-quarters of a minute to a minute. The Ach determinations made on a number of fishes whose skins had been prepared in this way were, however, quite diverse nor could any consistency in these differences be discovered. Since, moreover, the skins obtained by these steps did not appear uniformly pale, this technique was abandoned as probably too slow for the object at hand and a second and more rapid technique was tried.

In this second attempt fishes fully blanched were passed directly by hand from the tank water to alcohol chilled with dry ice to about $-20°C$. At this temperature the fish's skin was not only almost instantly frozen, but in somewhat less than 15 seconds after the fish entered the cold alcohol the fish itself was rigidly solid. Such a fish was then immersed in a large volume of water at 60°C. to insure the death of the skin after which the skin was removed, pulped, and extracted. This second method reduced the critical period during which the skin might darken to about a quarter of a

minute, and skins obtained in this way appeared to the experimenter's eye much more uniformly pale than did those made by the first technique. When, however, the skins prepared by this sudden and intense freezing were subjected to assays for Ach great diversity again reappeared in the results. The failure of both these attempts to obtain uniform Ach determinations from blanched catfish skins led to an abandonment of this type of technique and to the adoption of an entirely different approach to the problem at hand

This new line of attack depended upon the presence on the catfish of a considerable area of permanently pale skin in addition to that which is open to chromatic change. This area of pale skin is on the ventral aspect of the catfish and extends from the region of the opercular folds to that of the pelvic fins. In a fish somewhat over 15 cm. in length the pale area measures about 5 cm. long by some 3 cm. wide. It is white in tint and shows no change when the dorsal coloration of the fish alters under differences in the illuminated surroundings. It is of course not as ideal for this test as an area of blanched skin would be, but it appears to be the best substitute for blanched skin that under the circumstances can be found. When this ventral white skin is removed from a freshly killed fish, it remains permanently white. In making such preparations it is advisable to use catfishes in the dark phase. Under such circumstances two skin preparations can be obtained from each fish, a pale one from its venter and a dark one from its dorsum. Such preparations, in consequence of coming from the same fish, are especially favorable in the comparison of Ach determinations on the two kinds of skin. The detailed steps by which such determinations were made are given in the following description.

Dorsal and ventral areas of skin were removed from two or three fishes and the dark and the pale skins pooled separately. After blotting to remove excess of water, the tissue was placed in tared vessels containing 2 cc. of unbuffered, cold-blooded Ringer, with eserine sulfate 1:5000, and weighed. Normal HCl was added to pH 3 to 4. The skins and fluid were then heated for 3 minutes in a water bath at 90°C. This served the double purpose of aiding the release of bound Ach and greatly facilitating the breaking up of the skin in the grinding process. Additional Ringer was then added to give 1 cc. for each 500 mgm. of tissue. The skins were ground with silica, centrifuged and the clear supernatant fluid decanted for assay. Dilution of the extract with sea water at the time of assay adjusted the pH to within a range favorable to the test preparation.

The extracts of pale and of dark skins were compared with one another and matched against known concentrations of Ach for estimates of their Ach content. The isolated heart of *Venus mercenaria* was used in all tests. The values obtained, as given in table 1, are expressed as the equivalents of the free base in gamma per gram of moist skin.

The results of five separate sets of determinations, each done on a different test heart, are presented in table 1. The estimates for dark skin range from 0.02 to 0.08 gamma of Ach per gram of skin;† those for pale skin from 0.005 to 0.04 gamma per gram. Such variation does not indicate real differences in the Ach content of skins of different groups of fishes, but results rather from the varied response of different hearts to substances in addition to Ach in the extracts. There was evidence of the presence of an inhibitory substance other than Ach in the extracts and differences between hearts, in sensitivity to this substance, doubtless accounts for much of the variation in values for given samples of skin. The relative values of Ach in dark skin to Ach in pale skin in the several sets of experiments are without question significant. They reveal, on the average, over three times as much of this neurohumor in dorsal, melanophore-bearing, dark skin as in ventral, melanophore-free, pale skin.

TABLE 1

FIVE ASSAYS OF CATFISH SKINS, DORSAL DARK AND VENTRAL PALE, FOR ACH EXPRESSED IN FRACTIONS OF A GAMMA PER GRAM OF MOIST SKIN. IN EACH ASSAY TWO OR THREE CATFISHES WERE USED FROM EACH OF WHICH BOTH DARK SKIN AND PALE SKIN WERE TAKEN. ORGANIC INDICATOR: HEART, *Venus mercenaria*

DATES OF TESTS, 1948	ACH, GAMMA PER GRAM OF SKIN DARK SKIN	PALE SKIN	DARK SKIN TIMES PALE
July 8	0.08	0.04 –	2+
Nov. 15	0.05	0.015	3.3+
Nov. 16	0.02	0.005 –	4+
Nov. 24	0.03	0.01	3
Nov. 26	0.02	0.005	4
Averages	0.04	0.015	3.3

These two types of skin call for careful comparison. Both contain under autonomic control an ample vascular equipment and an abundant supply of mucous glands. These two systems are about equally developed in the dorsal and in the ventral skin. In this respect they are in strong contrast with the autonomically controlled melanophore system which is present in the dorsal skin but absent from the ventral. This distribution of autonomic effectors is fully consistent with the Ach determinations for these two types of skin as recorded in this paper. In the ventral skin where only vascular and glandular components are present the quantity of Ach averages some 0.015 gamma, whereas in the dorsal skin where there is a rich aggregation of melanophores in addition to the vascular and glandular outfits the quantity of Ach reaches on the average 0.04 gamma. Since the excess of Ach in the dorsal skin over that in the ventral, some 0.025 gamma, is thus associated in the dorsal skin with the presence of melanophores, it is natural to conclude that this excess has to do with the system of dark color-cells and that the residue of Ach in the dorsal skin, 0.015 gamma, like that

in the ventral skin is concerned with the other autonomic systems, vasomotor and glandular.

If the conclusion is correct that some 0.025 gamma of Ach per gram of skin is to be relegated to the melanophores, in what way, it may be asked, is this amount associated with these dark color-cells? When the dark skin of a catfish is removed preparatory to testing it for Ach, it is cut off the fish in such a way that it carries with it about a distal millimeter or more of each dispersing nerve-fibre with its numerous branches, chromatic terminals and associated melanophores. The relation of these chromatic terminals to the melanophores is not a simple one. Each melanophore, unlike, for instance, a single muscle-fibre, is innervated by branches from several different chromatic nerve-fibres whose terminals to the number of several scores or more surround the color-cell. Thus a melanophore lies immersed in an aggregation of chromatic terminals instead of being associated with only one efferent terminal as in the case of a muscle-fibre. Where, in this complex, is Ach located and how is it concerned with melanophore activities?

When a catfish darkens it assumes its deep tint as a result of two agencies, intermedine from the pituitary gland and Ach from its nerve supply. Intermedine when injected into a pale catfish can cause it to become fully dark by inducing the pigment in its melanophores to pass from full concentration to full dispersion. Injected Ach, on the other hand, can incite only about half this change in that it can cause the fish to shift its tint from full pale to an intermediate gray, approximately midway between full pale and full dark. If into a catfish rendered intermediate in tone by Ach an appropriate amount of intermedine is injected, the fish will become fully dark. This confirms the view that Ach is a half-darkening activator, so to speak, and intermedine a whole one. Hence after a catfish is hypophysectomized it becomes limited as a result of the loss of its source of intermedine to color changes between pale and intermediate gray. From this standpoint hypophysectomized catfish may be used in the study of the localization of Ach.

If a hypophysectomized catfish with one or more denervated bands cut in its tail is placed in a white-walled, illuminated tank, two changes will occur: The chromatic nerve-fibres and their terminals in the denervated band will degenerate fully in some ten days or so and the whole fish, including its tail and the caudal bands, will become pale. If now this pale fish is transferred to a black-walled, illuminated tank, the whole fish with the exception of its caudal bands will become in the course of a day intermediate gray. The band during and for some time after this change will remain pale. If into such a fish an appropriate quantity of Ach is injected, no change in tint in the fish as a whole will be seen, but the pale bands in its tail will slowly take on a deeper tint till they agree in shade with that of the rest of the tail.

From these observations pertinent conclusions may be drawn. Since the fish devoid of its pituitary gland is without a supply of intermedine such darkening as it is capable of must be due to Ach. As this Ach darkening occurs regularly in innervated areas but never in denervated ones (caudal bands), it is fair to conclude further that the Ach has as its source the nervous portion of the color complex and not the melanophore portion. This view is confirmed by the fact already stated that a pale band, inactive because of the degeneration of its chromatic nerves and terminals, can be darkened by an injection of Ach. Thus Ach introduced from the outside can take the place of the internal chromatic nerves. Hence this evidence appears to rule out the melanophore as a source for Ach and to relegate this material to the nervous portion of the chromatic complex.

This complex in such pieces of skin as have been used in these tests consists of innumerable distal pieces of many-branched chromatic nerve-fibres with a total length of about a millimeter each whose final branches terminate in slightly swollen knobs, the chromatic terminals. These terminals as already stated are in close proximity to the melanophores. Is Ach a product of this nervous mechanism as a whole or of one of its parts, the branches of the dispersing chromatic nerve-fibres or the terminals? No conclusive answer can be given to this question, but some evidence concerning it can be found in the exact relation of denervated dark areas to the locations of the cuts in the nerves whereby these dark areas are induced. Such conditions are well illustrated in certain relations shown by the ophthalmic nerve in the catfish. This nerve and its branches pass anteriorly over the dorsal wall of the orbital space in this fish to emerge dorsally through a thin layer of bone and spread out over the under surface of the skin in their distribution to the anterior dorsal region of the fish's head. For the test at hand these branches may be cut at the anterior limits of the orbit where they pass out of this cavity and through the bone to gain the inner surface of the skin. Shortly after these nerves have been cut the area of skin thus denervated on the fish's head will be easily distinguished by its dark tint and it is a noteworthy fact that this denervated, dark area does not begin where the nerve branches enter the skin but appears first several millimeters anterior to this region. In other words the chromatic nerve-fibres pass anteriorly through the skin on the fish's head some distance before they reach the darkened area. This relation indicates that the chromatic nerve-fibres of themselves are probably in no direct way concerned with the darkening of the skin. It is only after the region of their terminals has been reached that the darkening of the skin is to be seen. Hence this evidence indicates that the chromatic nerve-fibres in themselves have no direct concern with the darkening response, but that this response must be induced by the distal ends of these fibres, in other words, by their terminals. Although this evidence as already intimated cannot be said to be fully conclusive, it is not

without weight. It certainly suggests a consistent picture of the relations of the so-called chromatic terminals to the darkening process.

Assuming the correctness of this conclusion these terminals must be the parts immediately concerned with the 0.025 gamma of Ach which is proper to the melanophore apparatus in a single gram of dark skin. Whether this extremely small amount of Ach is all carried in the fish's lymph and other tissue fluids directly next the dark color-cells or whether it is in part in these fluids and in part in the chromatic terminals cannot be stated. In our opinion Ach is formed in the terminals where it is for the most part stored to be liberated in minute amounts through the action of chromatic-nerve impulses in the normal darkening of the fish or released as a whole from the terminals in the process of artificial extraction. We hold this opinion notwithstanding the fact that the total amount of Ach as measured by us is in itself extremely small and that the added reduction implied in the last statement places the functional amounts of Ach in the darkening of fishes at almost inconceivably small figures.

Summary.—1. The dark dorsal skin of the catfish Ameiurus containing three autonomically controlled systems (blood-vascular, mucous glands and melanophores) yielded from 0.02 to 0.08 gamma (average, 0.04) of Ach per gram of moist skin.

2. The pale ventral skin of the catfish containing only two autonomically controlled systems (blood-vascular and mucous glands) yielded from 0.005 to 0.04 gamma (average, 0.015) of Ach per gram of moist skin.

3. The difference between the averages of these two yields, 0.025 gamma of Ach, appears to be the amount of this neurohumor concerned with the activity of the melanophores.

4. This Ach is not produced by the melanophores. It probably does not originate in the dispersing chromatic nerve-fibres but rather in their terminals where it is very likely stored to be used from time to time as an agent to excite the dispersion of melanophore pigment.

5. Admitting this interpretation of melanophore activation by Ach it would appear that this operation calls for an amount of this neurohumor almost infinitesimally small.

* The expenses of this investigation were met by a grant from the Permanent Science Fund of the American Academy of Arts and Sciences. We wish to express here our obligations to the officers of this Fund for their generous help.

† It is to be noted that the Ach determinations already quoted from Chang, Hsieh and Lu (1939) and from Parker (1940), namely, 0.077 gamma and 0.078 gamma, respectively, per gram of dark skin, fall within the range of determinations for dark skin given in table 1.

Brown, G. L., and Feldberg, W., "The Acetylcholine Metabolism of a Sympathetic Ganglion," *Jour. Physiol.*, **88**, 265–283 (1936).

Chang, H. C., Hsieh, W., and Lu, Y. M., "Light-pituitary Reflex and Adrenergic-

cholinergic Sympathetic Nerve in a Teleost," *Proc. Soc. Exptl. Biol. Med.*, 40, 455–456 (1939).

MacIntosh, F. C., "Liberation of Acetylcholine by the Perfused Superior Cervical Ganglion," *Jour. Physiol.*, 94, 155–169 (1938)

Parker, G. H., "Color Changes in the Catfish Ameiurus in Relation to Neurohumors," *Jour. Exp. Zoöl.*, 69, 199–233 (1934).

Parker, G. H., "On the Neurohumors of the Color Changes in Catfishes and on Fats and Oils as Protective Agents for Such Substances," *Pro. Amer. Philos. Soc.*, 83, 379–409 (1940).

Prosser, C. L., "Acetylcholine and Nervous Inhibition in the Heart of Venus mercenaria," *Biol. Bull.*, 78, 92–102 (1940).

INADEQUACIES IN PRESENT KNOWLEDGE OF THE RELATION BETWEEN PHOTOSYNTHESIS AND THE O^{18} CONTENT OF ATMOSPHERIC OXYGEN

BY M. D. KAMEN AND H. A. BARKER

DIVISION OF PLANT NUTRITION, UNIVERSITY OF CALIFORNIA

Communicated November 7, 1944

Various lines or evidence strongly indicate that the oxygen of the atmosphere has been formed mainly by photosynthesis. If this conclusion is assumed to be correct, it follows that any specific relation between the substrates and products of photosynthesis that can be demonstrated in the laboratory should also be observable on a vastly larger scale in nature provided other interfering factors are not operative. One such specific relation with which we shall be concerned is that between the O^{18} content of water and of photosynthetic oxygen. The O^{18} content of the oxygen produced by the green alga *Chlorella pyrenoidosa* and by two land plants, sunflower and coleus, has been shown to be determined by the isotope content of the water and to be independent of that of the available carbon dioxide. Consequently, one might expect that in nature the O^{18} content of the oxygen of the atmosphere would be equal to that of the water in the oceans, since it is estimated that about four-fifths of the total photosynthesis occurs there and would be different from that of naturally occurring carbon dioxide and carbonates. Actually just the reverse is true; the O^{18} content of atmospheric oxygen is closer to that of carbonates than to that of ocean water, the O^{18} content being higher in the carbonates and atmospheric oxygen than in water.

A careful examination of the pertinent biological and physical evidence has led us to the conclusion that this apparent inconsistency does not constitute valid evidence against the photosynthetic theory of the origin of

atmospheric oxygen. It appears probable that an eventual explanation will be found in terms of isotope exchange reactions between molecular oxygen and water or other oxygen-containing compounds. However, the available data on isotope equilibria in water-oxygen systems cannot explain the observed relations. Therefore, a further study of such systems is desirable. Also additional information on the ultimate origin of oxygen produced by marine plants is needed to verify and extend the results obtained with fresh water and land plants.

A more detailed discussion of the points mentioned above is given in the following sections.

Geochemical Evidence for the Photosynthetic Origin of Atmospheric Oxygen.—In support of the assumption that atmospheric oxygen has been and is now being produced according to the well-known equation of photosynthesis

$$n\ CO_2 + n\ H_2O = n\ (CH_2O) + n\ O_2 \qquad (1)$$

we shall mention only two types of geochemical evidence. First, the estimated quantity of reduced carbon of biological origin on the earth's surface in the form of coal, oil, peat, humus, etc., corresponds approximately to the quantity of oxygen in the atmosphere after a correction has been made for oxygen removal by the weathering (oxidation) of igneous rocks and other similar reactions.[1] Second, the present rate of photosynthesis is sufficient to produce the quantity of oxygen in the atmosphere in a few thousand years, thereafter maintaining the atmosphere at its present oxygen level.[2] From this latter fact it may be concluded that whatever the origin of the atmosphere, the oxygen now present there must have been formed mainly by photosynthesis.

Basic Facts Concerning Oxygen Isotopes and Isotope Exchange Reactions.—Before discussing the experiments on the origin of oxygen in photosynthesis and the distribution of O^{18} in nature, it will be helpful to review briefly some basic facts concerning oxygen isotopes and isotope exchange reactions.

Natural oxygen contains three isotopes with mass numbers 16, 17 and 18, and percentage abundances of 99.76, 0.041 and 0.20, respectively.[3,4] In "enriched" samples of oxygen the increase in weight is usually credited to the O^{18} isotope entirely since the O^{17} contribution is relatively small due to its low abundance. In this paper the term "tracer oxygen" will be reserved for oxygen artificially enriched with O^{18}. The term "natural" oxygen will be understood to mean oxygen with the above mentioned isotope composition.

If two oxygen-containing molecules such as water and carbon dioxide are allowed to react until isotope equilibrium is reached, they do not in general contain the same proportions of isotopes. The relative abundance of isotopes in the two molecules may be expressed by a quantity known as the "enrichment factor," defined by the ratio $(N_1/N_2)/(n_1/n_2)$ where N_1 and

N_2 are the numbers of each isotope in one compound and n_1 and n_2 are the corresponding numbers in the other compound. Enrichment factors for several reactions which are involved in the present discussion are given in table 1, along with the corresponding equilibrium constants. The values have been calculated[5] on the basis of theoretical considerations; experimental verification[6] is available only for reaction (3).

Examination of table 1 shows that the enrichment factors for all the reactions here considered are larger than unity; this means that there is a higher proportion of oxygen O^{18} in the carbon dioxide or oxygen than in the water in isotopic equilibrium with it. Reactions (3) and (5) are of greatest importance for the present discussion. In the system CO_2 (g)–H_2O(1), (reaction (3)), at 25°C., the O^{18} enrichment in the gaseous carbon dioxide is such that water prepared from it is denser than the water in equilibrium with it by some ten gamma units.* The kinetics of the reaction are such that equilibrium can be attained in a relatively short time under suitable conditions. The reaction is slowest but still measurable in alkaline solutions (pH = 10) in which the carbon dioxide is mainly in the form of the carbonate ion. As the pH is lowered into the acid range (pH $\lesseqgtr$ 7), the rate increases rapidly.[7] In the equilibrium system O_2(g)–H_2O(1), (reaction (5)), at 25°, water prepared from the oxygen has a density excess of about 1 gamma unit over the water. Nothing appears to be known about the kinetics of this reaction, but the rate is presumably very slow.

TABLE 1

ENRICHMENT FACTORS AND EQUILIBRIUM CONSTANTS OF SOME OXYGEN EXCHANGE REACTIONS (AFTER UREY AND GREIFF)

REACTION	EQUILIBRIUM CONSTANT 0°	25°	ENRICHMENT FACTOR 0°	25°
(2) $CO_2{}^{16}(g) + H_2O^{18}(g) = CO_2{}^{18}(g) + H_2O^{16}(g)$	1.128	1.110	1.064	1.054
(3) $CO_2{}^{16}(g) + H_2O^{18}(l) = CO_2{}^{18}(g) + H_2O^{16}(l)$	1.097	1.080	1.047	1.039
(4) $O_2{}^{16}(g) + H_2O^{18}(g) = O_2{}^{18}(g) + H_2O^{16}(g)$	1.048	1.041	1.024	1.020
(5) $O_2{}^{16}(g) + H_2O^{18}(l) = O_2{}^{18}(g) + H_2O^{16}(l)$	1.020	1.012	1.010	1.006

For the study of photosynthesis it is also important to know the relative O^{18} contents of dissolved and gaseous carbon dioxide in equilibrium. At present there is no direct experimental evidence on this point. However, it may be inferred that carbon dioxide in the gaseous and dissolved states has about the same O^{18} content from the fact that calcareous rocks of marine origin have essentially the same isotope content as gaseous carbon dioxide in equilibrium with water.[8] Obviously direct experimental comparison of the O^{18} contents of dissolved and gaseous carbon dioxide would be most desirable; this could be done by a study of carbonate precipitated from equilibrated carbon dioxide–water systems.

An additional observation of importance must also be mentioned here. Water containing O^{18} has a slightly lower vapor pressure than water containing O^{16}. Therefore, during the evaporation of ordinary water, O^{18} tends to concentrate in the residual liquor.[9]

For determining the O^{18} content of a compound, two general methods are available; one involves measuring the density of water prepared from the oxygen and the other involves the direct determination of the relative numbers of oxygen atoms of different weights by means of a mass spectrometer. The former method is more sensitive, while the latter has the advantage of requiring only very small oxygen samples.

For measuring very small density differences, the so-called "submerged float" method provides a very satisfactory technique.[†] In this method a small glass or quartz bubble is immersed in the liquid, the density of which is to be measured. The temperature at which the float neither sinks nor rises is determined to about 0.0002°C. and is compared with the corresponding temperature for a solution of known density. From the temperature difference and a knowledge of the rate of density change with temperature, the relative density and therefore the relative O^{18} content of the sample can be calculated. With a temperature control of 0.0002°C., a density difference of about 0.5 parts in ten billion or .05 gamma unit should be detectable. Such sensitivities are far greater than warranted by the purification procedures employed which, in general, limit reproducibility of measurements to a few tenths of a gamma unit. In the experiments to which we shall refer in the following two sections, a density difference of 1 gamma unit can be regarded as significant. Details of the method are given in references cited below.[**]

The sensitivity of the mass spectrometer method of estimating O^{18} varies somewhat with the design of the instrument, being 0.5–1.0 gamma unit with the best models and ranging up to 5 or 10 gamma units density difference with the less sensitive models. This method, therefore, is not suitable for the study of the very small differences in O^{18} content such as exist, for example, in equilibrium systems of water and carbon dioxide, unless a very good instrument is available.

With the material presented above in mind, we can now proceed to consider the experiments on the origin of oxygen in photosynthesis and the known facts concerning the O^{18} distribution in nature.

Studies on the Rôle of Water as the Source of Photosynthetic Oxygen.—Since the substrates of photosynthesis are water and carbon dioxide, it is obvious that the evolved oxygen must be derived ultimately from one or both of these compounds. Various theoretical considerations, which need not be reviewed here, lead to the conclusion that the oxygen of photosynthesis probably comes from water and not from carbon dioxide.[10] The first experimental study of this hypothesis was undertaken in 1941 making

use of tracer oxygen.[11] Because a knowledge of the ultimate source of photosynthetic oxygen is of great importance in interpreting data on the distribution of oxygen isotopes in nature, the evidence bearing upon this hypothesis will be discussed in some detail.

Let us first examine the logic underlying experiments on the origin of photosynthetic oxygen. To demonstrate that oxygen originates from water, for example, it is necessary to "label" the oxygen in the water and show that it is transformed into molecular oxygen by the photosynthetic process. It is also necessary to show that when oxygen in carbon dioxide is labelled, no tracer oxygen appears as molecular oxygen in the gas phase. Since all photosynthetic products are derived from water or carbon dioxide, this demonstration would be sufficient to establish water as the ultimate source of oxygen. In carrying out such an experiment, one serious complication arises. If either the water or the carbon dioxide is labeled with tracer oxygen, the isotope exchange reaction (3) will immediately begin to redistribute the O^{18} and eventually isotope equilibrium will again be reached. Because of this exchange, the initial relatively large difference in isotope contents of water and carbon dioxide will be effective in labeling these molecules only when the rate of redistribution of O^{18} is small in comparison to the rate of photosynthesis. If the rate of exchange is relatively great, isotope equilibrium will be approached before the substrate molecules have had time to enter into the photosynthetic reaction and the isotope contents of the carbon dioxide and the evolved oxygen will necessarily approximate that of the initial water because the latter is present in vastly greater amount. It is therefore of great importance in experiments of this sort to provide proof that the rate of isotope exchange is too slow to invalidate the results.

In the experiments of Ruben, *et al.*, cell suspensions of *Chlorella pyrenoidosa* were allowed to photosynthesize in a carbonate-bicarbonate buffer solution having a pH of about 10 In some experiments the water was labelled with tracer oxygen, in others the carbonate was labeled. In both cases, the O^{18} content of the evolved oxygen was equal to that of the water within the limits of accuracy of the mass spectrometer measurements and differed widely from that of the carbon dioxide. As was to be expected from the kinetic studies already cited, a slow exchange of O^{18} was observed to occur between the carbon dioxide and the water. The overall rate of this exchange under the experimental conditions was slow enough to be neglected in comparison with the rate of photosynthesis. It was therefore concluded that the oxygen originated solely from water.

Such a conclusion is reasonable, but by no means certain since it depends on the unproven assumption that the isotope exchange is no more rapid inside the cells, and especially in the chloroplasts where photosynthesis actually occurs, than in the outside medium where it was measured.

It is quite possible and, indeed, even probable that this assumption is incorrect since the rate of isotope exchange increases rapidly with increasing hydrogen ion concentration, and it is known that many plant cells have an internal vacuolar pH $\leqq$ 6, which is maintained more or less constant despite wide variations of pH in the external medium.[12] It can be calculated that at pH 6, the randomization of O^{18} is rapid enough to invalidate the conclusion that carbon dioxide is not a source of oxygen.

In view of the existing uncertainty as to the internal pH of the *Chlorella* cell and the consequent uncertainty as to the rate of the randomization of O^{18} at the site of photosynthesis, these experiments do not provide proof of the rôle of water as the precursor of oxygen.

Recently, new and apparently unambiguous experiments[13,14] on the origin of photosynthetic oxygen have been performed which support the previous conclusion that oxygen arises wholly from water. The above-mentioned difficulties resulting from randomization of O^{18} were avoided by making use of the small difference between the O^{18} contents of water and carbon dioxide in equilibrium to trace the origin of the photosynthetic oxygen. This was made possible by recourse to the very sensitive submerged float method to analyze for O^{18}. In these experiments, *Chlorella* suspensions were allowed to photosynthesize in ordinary water equilibrated with carbon dioxide by means of the enzyme carbonic anhydrase from beef blood. The O^{18} content of the evolved oxygen was found to be nearly the same as that of the water and significantly lower than that of the carbon dioxide.. Experiments with land plants (sunflower and coleus) yielded similar results.

It seems reasonable to conclude that the rôle of water as the precursor of oxygen has been established in the case of one fresh water alga, *Chlorella pyrenoidosa*, and two land plants, sunflower and coleus. However, it must be remembered that most photosynthesis occurs in the oceans and hence marine organisms are much more important as large-scale producers of oxygen than are fresh water algae or land plants. It is therefore desirable that the above experimental results should be verified and extended by studies on a number of typical marine algae. There is no reason to believe that such studies will show any significant differences from those already reported. However, pending their completion, any general statement concerning the origin of photosynthetic oxygen can be accepted only with some reservation. Also there exists the possibility that during photosynthesis isotope equilibrium is established between the evolved oxygen and water. No evidence for such equilibration was obtained in the experiments of Ruben, *et al.* However, Dole and Jenks found that the O^{18} content of photosynthetic oxygen is approximately 1 gamma unit higher than that of the water. This suggests that equilibration may have occurred in these experiments. Obviously, the possibility that Chlorella catalyzes the ex-

change equilibrium between gaseous oxygen and water should be reinvestigated.

The Distribution of O^{18} in Nature.—The available information on the distribution of O^{18} in nature may be summarized as follows.[15]

(1) Oxygen in fresh water has the same O^{18} content as that in rocks and ores not containing carbonates.

(2) Sea water has a slightly higher O^{18} content than fresh water. The work of a number of investigators shows a density excess of 1.3–2.3 gamma units.[15,16] The increase in density is in good agreement with the value predicted from the experimental results obtained in isotopic fractionation of water during evaporation.

(3) Water prepared from the oxygen of carbonate rocks is denser than fresh water by some 8 gamma units.[8] This density excess is close to the value predicted on the basis that the carbonate is deposited from carbon dioxide in the sea water which is in equilibrium with atmospheric carbon dioxide and consequently enhanced in O^{18} according to reaction (3).

(4) Water prepared from atmospheric oxygen shows a density excess over fresh water of 6.6 gamma units.[8] It is here that the discrepancy arises between the physical data and the demonstration that photosynthetic organisms evolve oxygen with an O^{18} content very near that of water. 2 gamma units can be accounted for as due to the effect of evaporation discussed in (2) above. This leaves a 4.6 gamma unit discrepancy. On the basis of reaction (5), assuming a temperature intermediate between 0° and 25 °C. and the attainment of equilibrium, one can predict a further concentration of O^{18} in the gaseous oxygen corresponding to a density excess of about 1 gamma unit. This effect has been observed in the experiments of Dole and Jenks.[14] The residual 3.6 gamma units, sometimes known as the "Dole effect," remain to be explained.

As mentioned previously, the suggestion[15,17] that carbon dioxide enriched in O^{18} contributes the necessary heavy oxygen during photosynthesis is invalidated by all the experimental evidence so far obtained. There remain the following possibilities:

(*a*) The enrichment factor calculated for reaction (5) is in error. This is extremely doubtful since the physical constants required for the calculation of this equilibrium are accurately known. The theoretical values for this reaction should in any event be as reliable as those for reaction (3) which has been experimentally verified.

(*b*) Reaction (4), which has a more favorable enrichment factor than reaction (5), may be contributing to the O^{18} enhancement of atmospheric oxygen. However, at or near the surface of the earth where the oxygen studied was collected, reaction (5) must dictate the equilibrium finally reached, because the O^{18} content of the water involved would be determined by the large amount of liquid water in comparison with the water vapor.

As we have already remarked, reaction (5) fails by a large margin to account for the Dole effect.

(c) As previously mentioned, it is possible, though unlikely, that Chlorella and the two land plants so far studied are not typical of photosynthetic plants, particularly marine algae, with respect to the origin of their oxygen. This possibility must be investigated.

Although an ultimate explanation of the Dole effect may be found to depend upon either a biochemical or a physical effect, the latter possibility seems to us much more probable. In any event the existing inconsistency between the experimental evidence on the source of photosynthetic oxygen and the distribution of oxygen in nature cannot be regarded as valid evidence against the biological theory of the origin of atmospheric oxygen.

The authors wish to express their appreciation to Professors R. T. Birge, D. R. Hoagland and M. Dole, H. C. Urey and J. Franck for numerous valuable suggestions. We are also indebted to Professor Dole for permission to quote the results of unpublished work by him and G. Jenks.

[1] Goldschmidt, M., "Grundlagen der quant. Geochemie," *Fortschritte d. Mineral. Kristal. u. Petrographie*, **17**, 112–156 (1933).

[2] Gaffron, H., "Photosynthesis," *Encyclopedia Britannica*, 1944 Edit.

[3] Smythe, W. R., *Phys. Rev.*, **45**, 299 (1934).

[4] Murphey, B. F., *Phys. Rev.*, **59**, 320 (1941).

[5] Urey, H. C., and L. Grieff, *JACS*, **57**, 321 (1935).

[6] Weber, L. A., Wahl, M. H., and Urey, H. C., *Jour. Chem. Phys.*, **3**, 129 (1935).

* 1 gamma unit corresponds to a density difference of 1 part in 10^6.

[7] Mills, G. A., and Urey, H. C., *JACS*, **62**, 1019 (1940).

[8] This point is further discussed below in the section on the distribution of O^{18} in nature.

[9] Washburn, E. W., Smith, E. P., and Frandsen, M., *Bur. Standards Jour. Res.*, **11**, 453 (1933).

† A second method for measuring small density differences has been developed by H. E. Wirt, T. C. Thompson and C. L. Utterbeck, *J. Am. Chem. Soc.*, **57**, 400 (1935). It is based on the measurement by means of an electrical ultra-micrometer of the difference in height of balanced columns of two liquids whose densities are being compared.

** This elegant and precise method appears to have been first developed by T. W Richards and J. W. Shipley, *J. Am. Chem. Soc.*, **34**, 599 (1912).

[10] van Niel, C. B., *Adv. in Enzymology*, **1**, Interscience Publishers, Inc., N. Y., 263 (1941).

[11] Ruben, S., Randall, M., Kamen, M. D., and Hyde, J., *J. Am. Chem. Soc.*, **63**, 877 (1941).

[12] Hoagland, D. R., and Davis, A. R., *Jour. Gen. Physiol.*, **5**, 627 (1923); also see Small, J., *Hydrogen Ion Concentration in Plant Cells and Tissues*, Bornträger, Berlin (1929), pp. 93 and 305.

[13] Dole, M., private communication.

[14] Dole, M., and Jenks, G., *Science* **100**, 409 (1944).

[15] Birge, R. T., *Reports of Progress in Physics*, **8**, 90 (1941).

[16] Gilfillan, E. S., *J. Am. Chem. Soc.*, **56**, 406 (1934); Greene, C. H. and Voskuyl, R. J., *Ibid.*, **61**, 1342 (1939); Wirth, H. E., Thomson, T. G., and Utterbeck, C. L., *Ibid.*, **57**, 400 (1935).

[17] Greene, C. H., and Voskuyl, R. J., *J. Am. Chem. Soc.*, **61**, 1342 (1939).

*PRODUCTION OF STAPHYLOCOCCUS STRAINS RESISTANT TO VARIOUS CONCENTRATIONS OF PENICILLIN**

BY M. DEMEREC

CARNEGIE INSTITUTION, COLD SPRING HARBOR, NEW YORK

Communicated December 5, 1944

It is a well-established fact that strains of bacteria resistant to various sulfa drugs, as well as strains resistant to penicillin, may readily be obtained by growing bacteria on media containing increasingly higher concentrations of the respective chemicals. The purpose of this study was to make a quantitative survey of the origin of resistant bacteria, and to clarify the genetic aspect of the mechanism through which resistance is formed. A preliminary report summarizing the results obtained is given below.

Material and Method.—A strain of *Staphylococcus aureus* obtained from the Northern Regional Research Laboratory, Peoria, Illinois, carrying the N.R.R.L. number 313, was used in these experiments. This particular strain is employed by several laboratories for assaying penicillin. Before the experiment was started, a broth culture was prepared with bacteria from a single colony, and from this broth culture three agar slants were inoculated. These three stock cultures were kept in a refrigerator and served daily as the source of inoculum for all experiments. In a long series of experiments conducted over a considerable period of time, this procedure yields material that should be genetically more uniform than if the stock were maintained by consecutive transfers.

Penicillin was taken from a lot of sodium salt of penicillin prepared by E. R. Squibb and Sons, New York, which was packed in ampules containing 25,000 Oxford units each. The material of one ampule was dissolved in 10 cc. of phosphate buffer of pH 6, and kept in the refrigerator as a stock solution containing 2500 Oxford units of penicillin per cc. From this, other stock solutions containing 250 and 25 units per cc. were prepared under sterile conditions. Assays made at intervals indicated that the potency of penicillin in the stock solutions was not affected by storage.

The resistance of the bacteria to penicillin was determined by mixing them with an agar-nutrient medium to which the penicillin solution had been added, and plating the mixture in a Petri dish. Precautions were taken not to have the agar warmer than 45 degrees Centigrade.

In order to have a check of the potency of the penicillin, an assay was made for every experiment by diluting the solution used in that experiment to one Oxford unit per cc. and making a standard Oxford cup test. In this way any appreciable decrease in the potency of the penicillin solution would have been detected.

Reaction of Staphylococcus to Penicillin.—The strain of bacteria used in

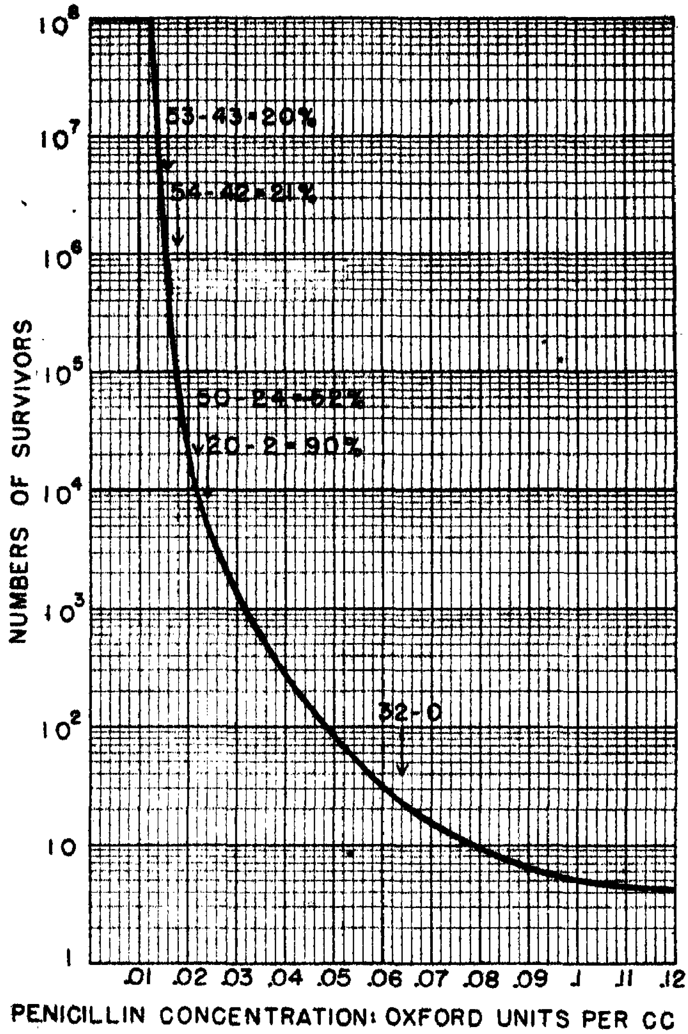


FIGURE 1

Numbers of surviving *Staphylococcus aureus* after plating on nutrient agar containing various concentrations of penicillin. Arrows indicate concentrations at which resistance to penicillin of surviving colonies was tested, and numbers above the arrows show the total number of colonies tested, the number of normal overlaps and the percentage of resistant colonies.

these experiments was affected by various concentrations of penicillin in a manner shown graphically in figure 1. No reduction in colony counts is apparent when the bacteria are plated onto nutrient-agar medium containing

weak concentrations of penicillin, until a threshhold of about 0.012 Oxford units per cc. is reached. A slight increase in concentration after this point produces a striking effect. For example, at a concentration of 0.014 only 10% of colonies appear; at a concentration of 0.016 the survival is 1%; at a concentration of 0.018, 1 per thousand; at a concentration of 0.05, about 1 per million; and a concentration of about 0.15 eliminates all bacteria. Six independent experiments gave very similar results; and the curve shown in figure 1 was fitted to the combined data of these experiments.

Thirty-two strains were established by inoculating broth with bacteria taken from 32 colonies which developed from surviving bacteria plated onto nutrient-agar medium containing 0.064 Oxford units of penicillin per cc. Not more than two colonies from any one plate were used for starting new strains, and the bacteria of the different plates were taken from different cultures. This precaution was taken in order to have strains that were not closely related. All of the 32 strains so established were tested for resistance to various concentrations of penicillin and were found to withstand concentrations considerably higher than the parent strain. The bacteria that formed colonies on plates with the nutrient medium containing 0.064 units per cc. of penicillin did so because they were resistant to at least that concentration of penicillin.

In a similar manner, 20 colonies which had formed on medium containing 0.024 units were tested on various concentrations of penicillin; 18 of them were more resistant than the original strain, while 2 were not. Therefore, at that concentration about 90% of surviving colonies had higher resistance than the parent strain, and 10% were normal overlaps—that is, chance survivors in a survival probability curve.

At a still lower concentration—namely 0.022 units per cc.—there were 50 survivors tested, 24 (or 48%) of which were normal overlaps. At a concentration of 0.018 units, 42 (or about 79%) of the 54 survivors tested were normal overlaps; and at 0.016 units, 43 (or 80%) of the 53 survivors tested were normal overlaps.

From these tests it is evident that at a concentration of 0.064 units per cc. of penicillin all tested survivors had higher resistance than the parent strain, while even at lower concentrations a considerable proportion of survivors were resistant. In figure 1 the arrows indicate concentrations from which survivors were tested, and the numbers above the arrows show the total number of tests, the number of normal overlaps and the percentage of resistant colonies.

Is the Resistance Inherited?—The evidence that has been accumulated suggests an affirmative answer to this question. A resistant strain isolated from the medium containing 0.064 units of penicillin was kept in the refrigerator on an agar slant for three months without any change in the de-

gree of resistance. Ten strains isolated in a similar manner were passed through 20 broth transfers, and tests for resistance were made at the end of that period as well as several times during the process. No change in the degree of resistance was observed. These strains, which were isolated after one passage in penicillin, acquired permanent resistance.

Origin of Resistant Bacteria.—Two alternate mechanisms can be visualized as responsible for the origin of bacteria resistant to certain concentrations of penicillin: (1) Resistance is an acquired characteristic, which develops through interaction between bacteria and penicillin when the two are in contact with each other. (2) Resistance is an inherited characteristic, which originates through mutation and whose origin is independent of penicillin treatment; resistant mutants occur at random, in a small fraction of a population, and, since a certain concentration of penicillin eliminates all non-resistant individuals, the resistant ones are selected out from the population by the treatment.

Which of these two mechanisms is responsible for the origin of resistance can be determined with the aid of a modification of the method developed by Luria and Delbrück[1] in their study of changes in bacteria from bacteriophage-sensitivity to bacteriophage-resistance. In the majority of experiments reported in this paper, bacteria were in contact with penicillin only during the time when the test for resistance was being carried on. Otherwise they were grown in the broth medium free of penicillin. If the resistance is induced through interaction between bacteria and penicillin when they are in contact with each other, it would be expected that approximately similar numbers of resistant bacteria would be obtained when samples containing similar numbers of bacteria are plated onto nutrient agar containing a certain concentration of penicillin, irrespective of the origin of these samples. The situation would be quite different in the event that the origin of resistance is mutational. In such case, one would expect to obtain similar numbers of resistant colonies only in samples taken from the same culture. If, however, each of the samples came from a separate culture, and mutations occur at random, then one would expect to obtain a large number of resistant colonies from cultures in which mutation happened to occur early in the growth of the culture and a small number of resistant colonies from cultures in which mutation happened to occur late, provided resistant bacteria grow more or less like the normal ones. If resistance originates by mutation, then, the variation in number of resistant bacteria between samples taken from separate cultures should be much greater than between samples taken from the same culture.

A critical experiment to distinguish between these two possibilities was planned as follows: A saturated broth culture of bacteria was diluted to 10^{-6}, so that the broth contained about 300 bacteria per cc. 0.3 cc. of this was placed in each of 30 small test tubes. One large test tube contain-

ing about 15 cc. of broth was inoculated with another sample of 0.3 cc. from the same dilution. All 31 test tubes were incubated at 37°C. for 18 hours, and precautions were taken to prevent evaporation from the tubes containing the small cultures. After 18 hours of incubation, 0.7 cc. of broth was added to each of these 30 small tubes, in order to reduce the error when the contents of each tube were taken out and plated. The number of bacteria was determined by sampling 10 of the small tubes, taking 0.05 cc. of material from each and assaying it on nutrient agar after making proper

TABLE 1

NUMBER OF BACTERIA RESISTANT TO CONCENTRATION OF 0.064 OXFORD UNITS OF PENICILLIN PER CC. OF AGAR MEDIUM IN SAMPLES TAKEN FROM A SERIES OF INDEPENDENT CULTURES AND SIMILAR SAMPLES TAKEN FROM A SINGLE CULTURE WHICH ASSAYED 2.3 × 10⁸ BACTERIA PER CC.

SAMPLES FROM INDEPENDENT CULTURES					SAMPLES FROM A SINGLE CULTURE	
CULTURE NO.	NO. OF BACTERIA PER CC.	NO. OF RESISTANT BACTERIA	CULTURE NO.	NO. OF RESISTANT BACTERIA	SAMPLE NO.	NO. OF RESISTANT BACTERIA
1	1.83×10^8	33	11	196	1	27
2	1.79	18	12	66	2	35
3	1.82	839	13	28	3	34
4	1.79	47	14	17	4	32
5	2.02	13	15	27	5	33
6	2.05	126	16	37	6	27
7	1.76	48	17	126	7	25
8	1.85	80	18	33	8	28
9	2.06	9	19	12	9	34
10	2.02	71	20	44	10	38
			21	28	11	25
			22	67	12	29
			23	730	13	31
			24	168	14	38
			25	44	15	31
			26	50	16	23
			27	583	17	16
			28	23	18	21
			29	17	19	30
			30	24	20	21
Average	1.9×10^8	128.4	...	116	..	28.9
Variance	1.35	57255	..	35399	..	39.8
χ^2	7.082	4459	..	6103	..	22.7
P	0.6	.....	..	...	..	0.3

serial dilutions. The number of bacteria in the large culture was determined by a similar assay. The material from the 30 small cultures (containing 0.3 cc. of saturated bacteria to which 0.7 cc. of broth had been added) was plated onto Petri dishes with nutrient agar containing 0.064 Oxford units of penicillin per cc. At the same time, twenty 0.3-cc. samples containing saturated bacteria from the large culture were plated in agar with 0.064 units of penicillin per cc.

Results of this experiment are given in table 1. It may be seen that the growth of bacteria in the different small cultures was fairly uniform, the average titre in ten cultures being 1.9×10^8 of individuals per cc.; and that this was fairly close to the titre of 2.3×10^8 per cc. reached in the large culture. That is, samples taken from the individual cultures and from the large culture contained approximately similar numbers of bacteria. It is evident from the table that the variation between the numbers of resistant bacteria on plates from samples taken from a single culture is small, the extremes being 16 and 38, the variance slightly larger than the average, and the probability that this variation is due to chance 30 per one hundred trials. On the other hand, the variation in number of resistant bacteria among samples taken from independent cultures is considerable, with extremes of 9 and 839 in cultures number 1 to number 10, and 12 and 730 in cultures numbers 11 to 20, with a variance greater than 200 times the average, and an insignificant probability that such a distribution may be due to sampling.

The results of this experiment, therefore, favor the assumption that resistance to certain concentrations of penicillin originates through mutation and that resistant bacteria may be found in any large population. The proportion of resistant bacteria depends on the mutation rate. The experiment was repeated three times, and similar results were obtained.

Another experiment furnished supporting evidence for the conclusion that resistance does not originate as a result of contact between bacteria and penicillin. In a study of the action of penicillin on Staphylococcus, it was found that penicillin affects principally dividing bacteria, while non-dividing bacteria can be kept for a considerable length of time in the penicillin-containing medium. To a saturated broth culture, containing 3.7×10^8 bacteria per cc., a sufficient amount of penicillin was added to make the concentration in the medium 25 Oxford units per cc. The culture was kept at 37°C., and after five days an assay showed that it contained 2.3×10^8 living bacteria per cc., while the assay for penicillin indicated that the concentration was not appreciably changed. These bacteria, exposed to 25 units of penicillin for five days, were washed to remove penicillin from the medium and tested for their resistance to various concentrations of this chemical. They were found to be no more resistant to penicillin than bacteria of the original strain. This shows that contact with penicillin does not make resting bacteria resistant.

Degree of Resistance.—In the strain of Staphylococcus used in these experiments, there is about one bacterium per 2×10^6 that survives a concentration of penicillin of 0.125 units per cc. At a concentration of 0.15 units per cc., there are no survivors. However, in strains established from survivors on an 0.064 concentration, there were a few individuals resistant to 0.15 units; in strains developed from survivors on an 0.125 concentration,

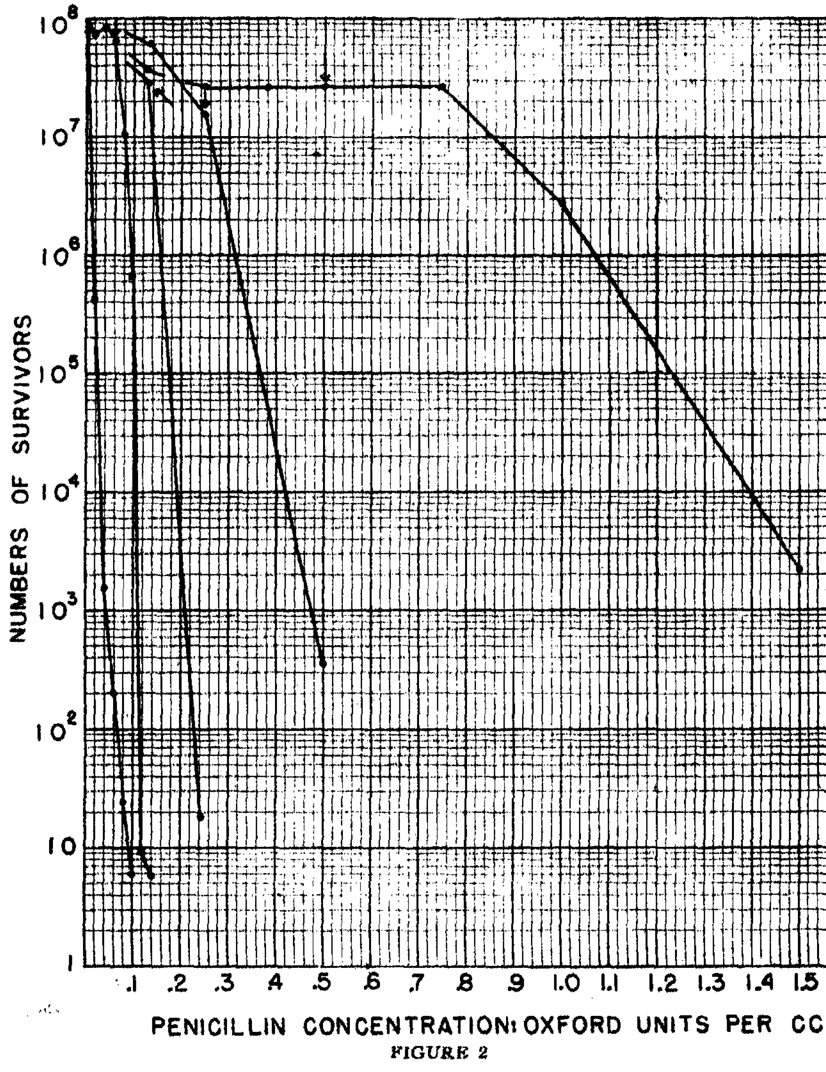


FIGURE 2

Numbers of survivors on various concentrations of penicillin, for the stock strain of *Staphylococcus aureus* and for four resistant strains developed through repeated selection.

there were individuals resistant to 0.25 units; strains from these latter survivors had individuals resistant to 0.5 units; strains from these contained individuals resistant to 4 units; and from these a strain was isolated that was not affected by a concentration of 250 units of penicillin per cc. of the agar medium.

The curves in figure 2 show the survival numbers, with various concentrations of penicillin, for bacteria of the stock strain and of four strains developed by the process of selection described in the previous paragraph. It is evident that the building up of resistance is more rapid with each selection step. That is, a concentration of 0.15 units is sufficient to eliminate all bacteria of the original strain, while to eliminate all bacteria of the first-step resistant strain a concentration of about 0.2 units is required, for the second-step resistant strain about 0.4 units, for the third-step about 1.0 units, and for the fourth-step about 7 units. The fifth-step strain was for all practical purposes completely resistant to penicillin.

As a check against contamination, a bacteriophage strain was isolated which lyses the Staphylococcus strain used in these experiments. This phage lysed also the completely resistant strain mentioned above, and all other resistant strains with which it was tested, indicating that these strains were derived from the original one rather than contaminants.

Discussion.—The evidence reported in this paper makes it probable that resistance of Staphylococcus to certain concentrations of penicillin is not induced by the action of penicillin on bacteria, but arises independently by mutation.

In any large population of bacteria of the strain of Staphylococcus used in these experiments there are some individuals resistant to certain low concentrations of penicillin. If this population is exposed to the action of such concentrations of penicillin, non-resistant individuals are eliminated while the resistant survive. Thus penicillin acts as a selective agent which suppresses non-resistant bacteria.

It is clear from the data that resistance is a complex characteristic, and that it must involve a number of mutations; if it is assumed that genes are responsible for these mutations, a number of genic changes must be involved. Such a situation is not unusual. A close parallel was described by Demerec and Fano[2] in the case of strain B of *Escherichia coli*, where about 20 distinguishable mutant types showing resistance to one or more of the seven phages were detected. Since it would be possible to isolate many more phages affecting the B strain of coli, it is evident that the actual number of mutants affecting resistance is considerably larger than the number detected.

It has been shown that degree of resistance can be increased by selection, and that the building up of resistance is more rapid with each selection step (Fig. 2). This can readily be explained by the mutation hypothesis. It is assumed that there are a number of genes that affect resistance to penicillin, and if any one of them mutates, the individual in which such mutation occurs acquires resistance to a certain concentration of penicillin. Mutants may differ in degree of resistance, but the resistance of individuals in which only one gene has mutated (single mutants) is never very high.

In a single-mutant strain, mutations may occur in other genes for resistance; and when two mutant genes are together in one individual (double mutants) their effect is cumulative. Moreover, it happens that the resistance of a double mutant is higher than the sum of resistances of two single mutants. If a third gene for resistance, a fourth, etc., mutate in the same line, the combined effect of all these mutations is a high degree of resistance or complete resistance.

Acknowledgments.—Acknowledgment is made of the very efficient assistance rendered by Miss Marion N. Crippen and Miss Carol Ruth McElroy in carrying out the experiments. The author is grateful to Dr. S. E. Luria for procuring from the Department of Surgery, College of Physicians and Surgeons, Columbia University, the bacteriophage active on the strain of Staphylococcus used in these experiments, as well as for helpful suggestions for improvements in bacteriological technique.

Summary.—In experiments with *Staphylococcus aureus*, strains resistant to penicillin were developed, which retained the property of resistance during the period covered by the experiments. Evidence is presented indicating that resistance is not induced by the action of penicillin on bacteria, but originates through mutation, and that penicillin acts as a selective agent to eliminate nonresistant individuals. Degree of resistance can be increased by exposure to higher concentrations of penicillin, and this increase is interpreted as due to summation of the effects of several independent genetic factors for resistance which undergo consecutive mutation.

* A portion of a research program which is supported in part by a grant from the Josiah Macy, Jr., Foundation.

[1] Luria, S. E., and Delbrück, M., *Genetics* **28**, 491–511 (1943).

[2] Demerec, M., and Fano, U., *Genetics* **30**, (in press) (1945).

THE LAW OF MASS ACTION IN EPIDEMIOLOGY

BY EDWIN B. WILSON AND JANE WORCESTER

HARVARD SCHOOL OF PUBLIC HEALTH

Communicated December 13, 1944

Almost all workers in the analytical theory of epidemics assume that the rate at which an infection passes in a population is proportional jointly to the product of the number of persons I who are infectious and the number of persons S who are susceptible to the infection.[1–6] This is called the law of mass action. Thus if the rate of new infections be C the law is written as

$$C = r\,I\,S, \tag{1}$$

where r is a constant. The *a priori* rationalization of the law generally is based upon the assumption, explicit or implicit, that the infectious I are mixing uniformly with the susceptibles S throughout the population. According to the law, if we had a population with twice as many susceptibles and infectious and with the same rate r of mixing, the rate C at which the infection passed would be not twice but four times as great. As a matter of fact, it is unlikely that any such condition exists in detail. For example, it is known that for the childhood infectious diseases such as measles the liability to infection within the family is greater than within the schoolroom and this is in turn greater than that within the community at large. The mixing of the susceptibles and infectious is not uniform throughout the population. Thus the real utility of the assumption for the explanation of the course of an epidemic must be found from the *a posteriori* observation that with the proper choice of a constant r the equation (1) yields a theoretical curve of new cases which is in satisfactory agreement with the observed curve of new cases. Such a value of r is presumably some complicated sort of average value of the different values of r under different degrees of intimacy of contact between different groups of infectious and susceptibles within the population

In the application of (1) the analytical developments vary according to the special assumptions made with respect to the particular disease under consideration. For example, if one is considering malaria and assumes that those once infectious remain so indefinitely and if one neglects accessions to or losses from the population and further neglects the incubation period whether in man or in mosquito, one writes

$$C = \frac{dI}{dt} = r\,I\,S, \qquad S = S_B - I, \tag{2}$$

where S_B is the number of susceptibles at the beginning; then (1) leads to

$$\frac{dI}{dt} = r\,I(S_B - I), \tag{3}$$

which on integration gives

$$I = {}^1/_2 S_B[1 + \tanh {}^1/_2 r S_B(t - t_0)], \tag{4}$$

$$C = {}^1/_4 r S_B^2 \operatorname{sech}^2 {}^1/_2 r S_B(t - t_0). \tag{5}$$

This means that the curve of total cases I is the logistic or growth curve, and the curve of new cases is symmetrical with respect to $t = t_0$. The rate of new cases when $t = t_0$ is $rS_B^2/4$ and the number of susceptibles remaining at that time is $S_B/2$, half of those at the beginning. In due time all the susceptibles are exhausted.

On the other hand, if the disease is one like measles in which it is generally assumed that there is an incubation period τ and a short period of infectiousness one may write

$$C = rSC(t - \tau) \text{ or } C = (S/m)C(t - \tau), \tag{6}$$

where $m = 1/r$ is the number of susceptibles just sufficient for one infectious case at $t - \tau$ to generate a new infectious case at t. Then, following Soper, and using $C = -dS/dt$ with $u = \log C$, one may obtain, to the order of approximation he uses,

$$\frac{d^2u}{dt^2} = -\frac{r}{\tau}e^u. \tag{7}$$

The integral is

$$u = 2 \log \operatorname{sech} \sqrt{\frac{rC_0}{2\tau}}\,(t - t_0) + \log C_0 \tag{8}$$

where C_0 is the rate of new cases when $t = t_0$, and then

$$C = C_0 \operatorname{sech}^2 \sqrt{\frac{rC_0}{2\tau}}\,(t - t_0). \tag{9}$$

It should be noted that the curve of new cases (or, more precisely, the curve of the rate of new cases) is under these assumptions and approximations of the same type as (5) which arose under very different assumptions.

In the third place if one modifies the law of mass action by assuming that the rate of new cases is proportional jointly to some power p of the number of susceptibles[7] and to the case rate lagged by τ, i.e.,

$$C = (S/m)^pC(t - \tau) \tag{10}$$

and eliminates S by $C = -dS/dt$ with $u = \log C$ one finds

$$\frac{d^2u}{dt^2} = -\frac{p}{m\tau}e^{u+(1/p-1/2)u'} \tag{11}$$

where the first correction term $(1/p - 1/2)(du/dt)$ has been kept in the exponent on the right. This term is neglected by Soper and was neglected, in the analysis above, for $p = 1$; it would appear to be equally negligible for other small values of p, and will be neglected. The integral is then

$$C = C_0 \operatorname{sech}^2 \sqrt{\frac{pC_0}{2m\tau}}\,(t - t_0), \tag{12}$$

and is still of the same type as (9) and (5). It is therefore clear that the form of the curve of new cases, apart from an interpretation of the constants which are involved and the assumptions which have been made in its derivation, cannot discriminate between a number of different laws of epidemic spread.[8] One could apparently get Soper's equation (7) and an epidemic curve of new cases as the derivative of the growth curve with better approximation from (11) if $p = 2$ than if $p = 1$.

As a matter of fact, one may show directly that when $p = 2$ equation (10) with $C = -dS/dt$ is exactly satisfied by a solution of type (12). For, given

$$\frac{dS}{dt} = \left(\frac{S}{m}\right)^2 \frac{dS}{dt}\bigg|_{t-\tau} \tag{10'}$$

we may substitute therein

$$S = m\left[\cosh\alpha - \sinh\alpha \tanh\alpha\,\frac{t - t_0}{\tau}\right] \tag{13}$$

and find that the equation is satisfied identically. Then

$$C - \frac{dS}{dt} = \frac{m\alpha}{\tau}\sinh\alpha\,\mathrm{sech}^2\,\alpha\,\frac{t - t_0}{\tau} \tag{14}$$

and the value of α is connected with the case rate when maximum by

$$C_0 = \frac{m}{\tau}\,\alpha\sinh\alpha = \frac{m}{\tau}\left[\alpha^2 + \frac{\alpha^4}{6} + \frac{\alpha^6}{120} + \ldots\right] \tag{15}$$

or

$$\alpha = \sqrt{\frac{C_0\tau}{m}}\left[1 - \frac{C_0\tau}{12m} + \frac{29}{1440}\left(\frac{C_0\tau}{m}\right)^2 - \ldots\right] \tag{15'}$$

For this case the initial and final values of S are

$$S_B = m(\cosh\alpha + \sinh\alpha), \qquad S_E = m(\cosh\alpha - \sinh\alpha)$$

and the value of S at the peak of the epidemic is $S_0 = m\cosh\alpha$ which is halfway between the initial and final values; moreover, $S_BS_E = m^2$ so that the "equilibrium value" m is the geometric mean of the initial and final values of S.

If we return to the general case where $p \neq 2$ and approximations are made in deriving (12) we obtain on integrating (12)

$$S = \text{const} - \sqrt{\frac{2m\tau C_0}{p}} \tanh \sqrt{\frac{pC_0}{2m\tau}}\,(t - t_0). \qquad (13')$$

The total number of cases from beginning to end of the epidemic is

$$\text{Total cases} = S_B - S_E = 2\sqrt{\frac{2m\tau C_0}{p}} \qquad (16)$$

and to the order of approximation used we find

$$\frac{m}{p} = \frac{(\text{total cases})^2}{8(\text{peak cases})}, \qquad (17)$$

provided we agree to call $C_0\tau$, which is the case rate at the peak of the epidemic multiplied by the incubation interval τ, the "peak cases." Thus what could be determined from an observed epidemic would not be either m or p severally but their ratio. In any such determination it would, of course, be necessary to use the estimated real numbers of total cases and of peak cases and not the total cases or peak cases reported unless the reporting were complete. It should further be observed that actually an epidemic may last over a considerable time and that recruits are coming into the population of susceptibles, which might well make necessary some modification in (17)

Instead of pursuing these considerations at this time we shall turn to the matter of the exact stepwise integration of (10). For notational simplification we introduce as in earlier papers,[8] $x = S/m$, and $T = t/\tau$ so that (10) becomes

$$\frac{dx}{dT} = x^p \left.\frac{dx}{dT}\right|_{-1} \text{ or } \frac{1}{x^p}\frac{dx}{dT} = \left.\frac{dx}{dT}\right|_{-1}. \qquad (10'')$$

The equation may be integrated exactly as[9]

$$\frac{1}{qx_T^q} + x_{T-1} = k = \frac{1}{qx_0^q} + x_{-1}, \qquad (18)$$

where $q = p - 1$ and x_0, x_{-1} are any two values of x which are one incubation period apart. At the beginning and end of the epidemic there are no cases and $x_T = x_{T-1}$. Hence the equation

$$\frac{1}{qx^q} + x - k = 0 \text{ or } \frac{1}{q}e^{-qv} + e^v - k = 0 \qquad (19)$$

with $v = \log x$ will have as solutions the initial and final values x_B, x_E of x or their logarithms. We have

$$f(v) = \frac{1}{q} e^{-qv} + e^{v} - k, \qquad f'(v) = -e^{-qv} + e^{v},$$

$$f'(v) + f'(-v) = 2(\cosh v - \cosh qv).$$

Hence the plot of $f(v)$ has a minimum at $v = 0$, and the (positive) slope for a positive value of v is numerically greater than the (negative) slope for the same numerical but negative value of v, provided $q < 1$, but for $q > 1$ it is less. This means that $v_B + v_E < 0$ or $x_B x_E < 1$ when $q < 1$, i.e., when $0 < p = 2$, but that $x_B x_E > 1$ when $q > 1$, i.e., when $p > 2$.

At the end of the epidemic where the case rates are very small and the values of x are not changing appreciably, C/C_{-1} being x^p is essentially constant and hence the curve of case rates in portions remote from the mode is essentially an exponential curve with a constant difference $p \log x_B$ for the ascending tail and with a constant (negative) difference $p \log x_E$ for the descending tail. The rise will be faster than the fall if $x_B x_E > 1$, i.e., if $p > 2$, but will be slower than the fall if $x_B x_E < 1$, i.e., if $p < 2$. We have seen that for $p = 2$ the curve of case rates is strictly symmetrical. It appears, however, that for $p > 2$ the longer tail would be on the right whereas for $p < 2$ it would be on the left.[10]

[1] Ross, Sir Ronald, "Application of the Theory of Probabilities to the Study of *a priori* Pathometry," Part 1, *Proc. Roy. Soc. London*, **A92**, 204–230 (1915); Ross, Sir Ronald, and Hudson, Hilda P., *Idem.*, Parts 2–3, *Ibid.*, **A93**, 212–240 (1917). The treatment is very general and not limited to the study of epidemics; the law of mass action is introduced under the term "proportional happening"; Part 1, pp. 220 ff.

[2] Lotka, A. J., "Contribution to the Analysis of Malaria Epidemiology," Supplement to *Amer. Jour. Hygiene*, **3**, 1–121 (1923).

[3] Soper, H. E., "The Interpretation of Periodicity in Disease Prevalence," *Jour. Roy. Statist. Soc. London*, **92**, 34–73 (1929).

[4] Frost, W. H., Cutter Lectures, Harvard Medical School, Feb. 2–3, 1928 (unpublished). The method, somewhat adapted, was used in Zinsser, H. and Wilson, E. B., "Bacterial Dissociation and a Theory of the Rise and Decline of Epidemic Waves," *Jour. Prev. Med.*, **6**, 497–514 (1932).

[5] Kermack, W. O., and McKendrick, A. G., "Contributions to the Mathematical Theory of Epidemics," *Proc. Roy. Soc., London*, **115**, 700–721 (1927); **138**, 55–83 (1932); and McKendrick, A. G., "The Dynamics of Crowd Infections," *Edinburgh Med. Jour.*, **47**, 117–136 (1940), where other references are given.

[6] Wilson, E. B., and Burke, M. H., these PROCEEDINGS, **28**, 361–367 (1942); **29**, 43–48 (1943); and Wilson, E. B., and Worcester, J., *Ibid.*, **30**, 37–44 and 264–269 (1944).

[7] A word should be said about the assumption that we might use some power p of S in (10). The assumption may be difficult to justify on *a priori* grounds, but the justification for the case $p = 1$ is none too satisfactory. It would, in fact, be remarkable in a situation so complex as that of the passage of an epidemic over a community if any simple law adequately represented the phenomenon in detail—even to assume that the new case rate should be set equal to any function $f(S)$ of the susceptibles multiplied by the case rate one incubation period earlier might be questioned. We propose to discuss the assumption (10) merely as a possible empirical variant of the case $p = 1$ to see what its

consequences may be. Although mathematics is used to develop the logical inferences from known laws, it may also be used to investigate the consequences of various assumptions when the laws are not known, i.e., one of the functions of mathematical and philosophical reasoning is to keep us alive to what may be only possibilities when the actualities are not yet known.

[8] Ross (Part I, p. 226) develops under the case of "proportional happening" considerations of circumstances which he says "are probably just the conditions which hold in many of the short and sharp epidemics of zymotic diseases, such as measles, scarlatina and dengue." It is perfectly true that the curve of new cases has the form found in epidemics of those diseases, but we have seen that this sort of curve may arise under a variety of different hypotheses. It seems tolerably clear that Ross's theory of happenings, despite its generality, does not include the hypotheses appropriate to the discussion of epidemics of such diseases as measles, for he assumes that his population P has only two divisions, namely, the affected population Z and the susceptible population A and that immunity and the affected condition disappear together. In the case of measles and similar diseases there is a third population, namely, the immunes, let us say Y, so that $P = Z + A + Y$ and his fundamental equations would be replaced by something like:

$$
\begin{aligned}
dP &= (n - m + i - e)Adt + (N - M + I - E)Zdt + (N' - M' + I' - E')Ydt, \\
dA &= (n - m + i - e - h)Adt + (N + r)Zdt + (N' + r')Ydt, \\
dZ &= hAdt + (-M + I - E - r - s)Zdt, \\
dY &= (-M' + I' - E' - r')Ydt + sZdt.
\end{aligned}
$$

Here n, m, i, e and their correlatives in capitals are natality, mortality, immigration and emigration rates, r is the rate at which the affected return to the susceptibles directly according to Ross's assumption of simultaneous cure and loss of immunity, r' is the rate at which the immunes lose their immunity, s is the rate at which the affected become cured and immune, and h is a factor which under the assumption of proportional happening has the form cZ. These equations do not allow for lag; they assume that births to the affected and to the immunes are susceptible rather than either affected or immune, though for measles children born to the immune mothers are generally themselves immune for a time.

We have tried to reconcile Ross's formulation (which is abstracted by C. O. Stallybrass in his *Principles of Epidemiology*, 1931, pp. 515 ff) and in particular the statement that in infectious diseases the reversion element $rZdt$ implies loss of both immunity and infectiousness, not recovery from disease, by considering the population of affected persons Z to remain affected whether infectious or not as long as they remain immune, but this construction appears impossible. We therefore seem to be forced to the conclusions that Ross's *a priori* pathometry does not cover those zymotic diseases in which immunity with non-infectiousness is a prime phenomenon; it might well cover those in which immunes were permanent carriers, especially if the rates of transfer of infection from the ill and from carriers to susceptibles were not materially different.

[9] If in equation (18) we take $x = 1$ when $T = 0$ and assume $x = 1 + aT + bT^2 + cT^3 + dT^4 + eT^5$, we find, on equating coefficients of powers of T, the following values of b, c, d, e, k in terms of a, good to the power a^5, inclusive.

$$
b = pa^2\left[\frac{1}{4} + \frac{p-2}{48}a + \frac{p^2-2p}{192}a^2 + \frac{p^3-2p^2}{768}a^3\right]
$$

$$c = pa^2\left[\frac{1}{6} + \frac{7p-6}{72}a + \frac{11p^2 - 21p + 12}{432}a^2 + \frac{97p^3 - 222p^2 + 168p}{10368}a^3\right]$$

$$d = pa^3\left[\frac{5p-2}{48} + \frac{8p^2 - 15p + 12}{288}a + \frac{79p^3 - 186p^2 + 168p}{6912}a^2\right]$$

$$e = pa^3\left[\frac{p}{30} + \frac{13p^2 - 24p + 24}{1440}a + \frac{35p^3 - 84p^2 + 84p}{8640}a^2\right]$$

$$k = \frac{p}{p-1} - a + \frac{p}{12}a^2 - \frac{p^2}{180}a^3 - \frac{13p^3 - 24p^2 + 24p}{8640}a^4 - \frac{35p^4 - 84p^3 + 84p^2}{51840}a^5.$$

For any assumed slope of the x-curve at $x = 1$ one could plot a short range of that curve, say from $T = -1/2$ to $T = +1/2$. With the value of k one could then proceed stepwise from (18) to any other value of x removed an integral number of units of time for any value assumed within that range.

Particularly interesting, however, is to derive expressions in terms of the maximum case rate $z_0 = C_0\tau/m$. This requires the value of T when $d^2x/dT^2 = 0$, which is

$$T_0 = -\frac{1}{2} + \frac{p-2}{48}a + \frac{5p^2 - 10p}{576} + \frac{11p^3 - 26p^2 + 12p - 8}{3072}a^3,$$

and is valid only to the term in a^3 whereas x was valid to the term in a^5. For this value of T, $z = -dx/dT$ takes the value z_0, viz.,

$$z_0 = -a + \frac{p}{8}a^2 - \frac{p^2}{96}a^3 \quad \text{or} \quad a = -z_0 + \frac{p}{8}z_0^2 - \frac{p^2}{48}z_0^3.$$

Hence a may be found in terms of z_0, good to the term in z_0^3, inclusive. With this value of a one may derive expressions in z_0 for k and for the value of x_0 of x at the peak of the epidemic, as follows:

$$k = \frac{p}{p-1} + z_0 - \frac{p}{24}z_0^2 + \frac{p^2}{180}z_0^3, \tag{A}$$

$$x_0 = 1 + \frac{1}{2}z_0 - \frac{1}{24}z_0^2 - \frac{113p^2 - 290p}{11520}z_0^3. \tag{B}$$

With these values one may compute stepwise from (18) the values of x for successive values of x removed from the mode by integral numbers of units of time under any assumed value of the ratio $C_0\tau/m$ and for any value of p.

For the approximation that leads to the symmetrical curve, we may write (13′) nearly enough as

$$S = m + \frac{1}{2}C_0\tau - \sqrt{\frac{2\tau m C_0}{p}}\tanh\sqrt{\frac{pC_0}{2\tau m}}(t - t_0). \tag{C}$$

Therefore for the beginning and end of the epidemic we should have

$$S_B = m + \frac{1}{2}C_{0\tau} + \sqrt{\frac{2\tau m C_0}{p}}, \qquad S_E = m + \frac{1}{2}C_{0\tau} - \sqrt{\frac{2\tau m C_0}{p}}.$$

In general for x at the beginning and end, equation (19) for x may be solved in series. If we set

$$Y = \frac{2}{p}\left(k - \frac{p}{p-1}\right) = \frac{2}{p}z_0 - \frac{1}{12}z_0^2 + \frac{p}{90}z_0^3,$$

$$x = 1 \pm Y^{1/2} + \frac{p+1}{6}Y \pm \frac{(p+1)(2p-1)}{72}Y^{3/2} + \frac{(p+1)(2p-1)(p-2)}{540}Y^2 \pm \frac{(p+1)(2p-1)(2p^2-23p+23)}{17280}Y^{5/2}$$

and x_B will be the value with the positive sign, x_E that with the negative, and

$$\text{Total cases} = 2Y^{1/2} + \frac{(p+1)(2p-1)}{36}Y^{3/2} + \frac{(p+1)(2p-1)(2p^2-23p+23)}{8640}Y^{5/2}$$

provided we measure cases relative to m. Transformed to actual numbers

$$\text{Total cases} = m\left[2Y^{1/2} + \frac{(p+1)(2p-1)}{36}Y^{3/2} + \frac{(p+1)(2p-1)(2p^2-23p+23)}{8640}Y^{5/2}\right], \quad \text{(D)}$$

$$Y = \frac{2}{p}\frac{C_{0\tau}}{m} - \frac{1}{12}\left(\frac{C_{0\tau}}{m}\right)^2 + \frac{p}{90}\left(\frac{C_{0\tau}}{m}\right)^3. \quad \text{(E)}$$

Equation (D) with Y defined as in (E) gives a relationship between total cases, peak cases $C_{0\tau}$ and m and p which may be solved for any one of those four quantities in terms of the other three to give a more exact expression than (17) which was based on an approximation. If we solve for m and retain only the first approximation beyond (17) we find

$$m = p\frac{(\text{total cases})^2}{8(\text{peak cases})}\left[1 - \frac{5p^2+4p-4}{9p^2}\left(\frac{\text{peak cases}}{\text{total cases}}\right)^2\right]. \quad \text{(F)}$$

For $p \geqq 1$ the correction term in (F) is at most two-thirds of the square of the ratio of peak cases to total cases. This ratio in sharp epidemics of measles (after allowance for under-reporting) is rarely as much as $1/6$ so that the correction rarely amounts to more than two percent, and, therefore, considering the difficulty of accurate estimation of actual total cases or peak cases, we may consider (17) a sufficiently good approximation.

[10] The actual calculation of the course of $S = mx$ from (18) for three comparative special cases, with $z = -dx/dT = C/m$, is as follows:

$p = 1, m = 1, z_0 = 0.3$ S	Cases	$p = 2, m = 2, z_0 = 15$ S	Cases	$p = 3, m = 3, z_0 = 10$ S	Cases
1.9789					
	0.0001				
1.9788		2.9326			
	0.0002		0.0002		
1.9786		2.9324			
	0.0004		0.0002		
1.9782		2.9322		3.9165	
	0.0008		0.0004		0.0003
1.9774		2.9318		3.9162	
	0.0016		0.0008		0.0006
1.9758		2.9310		3.9156	
	0.0032		0.0018		0.0015
1.9726		2.9292		3.9141	
	0.0063		0.0038		0.0033
1.9663		2.9254		3.9108	
	0.0123		0.0084		0.0072
1.9540		2.9170		3.9036	
	0.0238		0.0178		0.0156
1.9302		2.8992		3.8880	
	0.0455		0.0368		0.0336
1.8847		2.8624		3.8544	
	0.0837		0.0732		0.0694
1.8010		2.7892		3.7848	
	0.1446		0.1356		0.1320
1.6564		2.6536		3.6528	
	0.2231		0.2190		0.2175
1.4833		2.4346		3.4353	
	0.2866		0.2864		0.2865
1.1467		2.1482		3.1488	
	0.2857		0.2864		0.2865
0.8610		1.8618		2.8623	
	0.2140		0.2190		0.2205
0.6470		1.6428		2.6418	
	0.1246		0.1356		0.1389
0.5224		1.5072		2.5029	
	0.0612		0.0732		0.0768
0.4612		1.4340		2.4261	
	0.0274		0.0366		0.0396
0.4338		1.3974		2.3865	
	0.0117		0.0178		0.0198
0.4221		1.3796		2.3667	
	0.0049		0.0084		0.0096
0.4172		1.3712		2.3571	
	0.0021		0.0038		0.0048
0.4151		1.3674		2.3523	
	0.0009		0.0018		0.0024
0.4142		1.3656		2.3499	
	0.0003		0.0008		0.0009
0.4139		1.3648		2.3490	
	0.0002		0.0004		0.0006
0.4137		1.3644		2.3484	
			0.0002		0.0003
		1.3642		2.3481	
			0.0002		
		1.3640			

The left-hand skewness for $p = 1$ and right-hand skewness for $p = 3$ show in the figures. The different values of z_0 were chosen so that the case rates at maximum $C_{0T} = z_0 m$ and total cases should be the same provided the approximate formula (16) were used and the value of x_0 were taken from (B) in footnote 9. Slight irregularities in the numbers must be expected due to the limited number of places carried, and slight discrepancies in verifying (16) from the calculations because of the approximative nature of (16) and (B).

J. Brownlee stated, *Proc. Roy. Soc. Med., Epid. Sect.*, **2**, Part 2, 243–258 (1909): . . . the symmetry of the course of the epidemic is an obvious and marked feature. The deduction from this phenomenon is direct and complete, namely, that the want of persons liable to infection is not the cause of the decay of the epidemic. On no law of infection which I have been able to devise would such a cause permit epidemic symmetry. The fall must in all cases be much more rapid than the rise, though, on the contrary, when asymmetry is markedly present the opposite holds. Ross[1] comments on this statement. We may point out that if we accept the generalization of the law of mass action suggested in (10) there is symmetry for $p = 2$, negative skewness for $p < 2$ and positive skewness for $p > 2$. Thus a rather simple law has been devised which may explain symmetry or skewness of either sign. Furthermore, in the examples above which correspond to rather severe epidemics of measles the rise is at the (logarithmic) rate $p \log x_B$ or 0.68, 0.77, 0.81, respectively, for $p = 1, 2, 3$; and the rate of the fall is $-p \log x_E$ or 0.89, 0.77, 0.74, respectively. In the first case the rate of fall is considerably greater than the rate of rise, in the second case they are equal, and in the third case the rate of fall is but slightly less than the rate of rise. With higher values of p the rate of fall would become considerably less than the rate of rise, but even with very high values of p and with the same values of peak cases and of total cases as in the illustrations above the rate of rise could probably not exceed 0.89 and the rate of fall not be lower than 0.69.

A LETTER FROM LORD RAYLEIGH TO J. WILLARD GIBBS AND HIS REPLY

BY EDWIN B. WILSON

HARVARD SCHOOL OF PUBLIC HEALTH

Communicated December 4, 1944

In the small collection of letters left by J. W. Gibbs and now in the possession of Ralph G. Van Name is one from Lord Rayleigh the answer to which I presumed still existed because the present Lord Rayleigh quoted three sentences from it in his biography of his father.[1] When I sent a copy of his father's letter to Lord Rayleigh, he kindly sent me a transcript of Gibbs's reply. As this exchange of letters between a foreign associate and a member of this Academy seems to me likely to be of sufficient interest to our members and of sufficient importance to the history of science to justify publication in full even at this late date, I have secured the permission of Lord Rayleigh and of Professor Van Name to print them here.

June 5/92
Terling Place,
Witham, Essex.

Dear Prof. Gibbs:

I have been experimenting lately upon the intensity of reflection from water at nearly perpendicular incidence, and the not very close agreement that I have found with Young's formula $\left(\frac{\mu - 1}{\mu + 1}\right)^2$ has set me thinking whether there is any reason for expecting this formula to be correct, when the effect of *dispersion is included.* The case of two strings joined together one (as usual) perfectly flexible and the other with such stiffness as to introduce dispersion, shews that the reflection is *not* to be got from $\frac{\mu - 1}{\mu + 1}$, even though the correct μ be used; I think the deviation from this is of the same order as the deviation of $\frac{\mu_\lambda - 1}{\mu_\lambda + 1}$ from $\frac{\mu_\infty - 1}{\mu_\infty + 1}$. But this case does not much resemble optical dispersion (one would suppose), and I am writing partly to ask whether you have ever considered the problem of reflection with inclusion of dispersion on the lines of your published papers. And with respect to the latter I find a difficulty in your estimation on Kelvin's theory [is not this analytically identical with Lorenz?] of the potential energy as

$$\frac{\pi^2 Bh^2}{l^2} + \frac{bh^2}{4}$$

(*Phil. Mag.*, XXVII, p. 24) in which the 1st contains l and the 2nd does not. Supposing the disturbance from uniformity to be by simple changes of elastic quality, would not the 2nd term contain l^2 like the first.

And now on another subject. Have you ever thought of bringing out a new edition of, or a treatise founded upon, your "Equilibrium of Het. Substances." The original version though now attracting the attention it deserves, is too condensed and too difficult for most, I might say all, readers. The result is that as has happened to myself, the idea is not grasped until the subject has come up in one's own mind more or less independently. I am sure that there is no one who could write a book on Thermodynamics like yourself.

I feel that I am taking a liberty in writing like this, but it is in the interest of science and you will forgive me.

I remain
yrs very truly
Rayleigh

New Haven.
June 27, 1892.

My dear Lord Rayleigh,

The electrical theory of light seems to afford a very simple equation for harmonic motions of any one period in any optical field, *without* neglect of the causes of dispersion. If we consider an element of space containing

many ponderable molecules, the electrical motions in that space are presumably very complicated, but they result from the motions at a distance, and the effect of motions at a distance can be expressed by a very simple formula. There is first the electromotive force of induction, of which the components are calculated from the components of acceleration by the same law as the potential is calculated in the theory of gravitation from the density of matter. If we write $\overline{\text{Pot}}$ to express this operation, we will have $-\text{Pot}\,\ddot{F}$ for this force, F being a bi-vector representing the displacement (*i. e.*, the three components of F are complex realms). There is also (or may be) an electrostatic force which can be represented by ∇Q, where Q is a complex realm (the electrostatic potential). It should be borne in mind that the displacements which occur in the equations of wave motion are a sort of average for elements of space which are large in comparison with the distances of neighboring molecules, but small in comparison with a wave-length. If F is understood as representing such an average, and likewise Q

$$-\text{Pot}\,\ddot{F} - \nabla Q$$

will still correctly represent the electromotive forces acting upon the element from a distance. Whether this formula correctly represents the internal forces of the element or those forces which arise from the immediate neighbourhood of the element is of no consequence. Now I say that F is a function of $-\text{Pot}\,\ddot{F} - \nabla Q$. This is really saying very little. It is only saying that if under any circumstance an average harmonic motion F subsists in an element of space, and if we then change things in the remoter parts of the field, but so that the value of $-\text{Pot}\,\ddot{F} - \nabla Q$ shall remain unaltered at the element considered, the same motion F will continue to subsist in that element. F is therefore a function of $-\text{Pot}\,\ddot{F} - \nabla Q$ and by the principle of superposition of motions a linear function. Or, we may say that $-\text{Pot}\,\ddot{F} - \nabla Q$ is a linear function of F, and write as in my paper on Kelvin's quasi-labile ether.

$$-\text{Pot}\,\ddot{F} - \nabla Q = 4\pi\Phi F$$

the operator Φ representing in general a linear vector function, which however in the case of an isotropic body reduces to a (so-called) numerical quantity (*real*, if the body is transparent). It will of course vary in different parts of the field with the optical properties of the bodies in the field. The equation is not indeed absolutely accurate—the definitions on which it is founded are not absolutely sharp, and the phenomena of (non-magnetic) rotation of plane of polarization form a striking exception to the equation, which I have discussed at length in my 2nd "Note on the Elec. Theory of Light." We may say that the inaccuracy in the equation results from the virtual assumption that the structure of a body is infinitely fine as measured by a wave-length of light.

But I do not see that any amount of *dispersion* constitutes any reason for the failure of the equation. It will apply, so far as I can see, even to cases of selective absorptive and abnormal (so-called) dispersion, as well as to any degree of opacity.

Now this general equation (as I have pointed out in my paper on Kelvin's

quasi-labile ether) gives us on one hand Fresnel's wave-surface, and on the other Fresnel's laws of intensities of reflected light, if we suppose that Φ is constant on each side of a mathematical plane of reflection, which is of course a very precarious supposition. You will remember that in one of your papers you refer to one of Lorentz, in which he claims to obtain a remarkable agreement with Jamin's experiments for solids by supposing the change in index (i.e., in Φ) to be gradual. Your own experiments have shown that at least in the case of liquids the influence of foreign substances at the surface is considerable. Besides the question of the gradual or abrupt change in Φ there is the question (at least when the foreign substances are present) of values of Φ which are not intermediate, as complex values where Φ is real on each side of the surface.

I was very glad that you vindicated Fresnel's law from the most considerable deviation as reported by Jamin for liquids, and feel quite sorry that your later experiments show in another way a deviation. I should hope for the sake of the theory (to which you see I am somewhat attached) that the deviations can be accounted for by attributing some not unnatural qualities to a thin film at the plane of reflection.[2]

With respect to reflection of waves in strings, we may cause dispersion by supporting a heavy string by threads (very close together) from the ceiling. For waves of any one period (in time) the reflection where two such strings are joined is determined entirely by the wave-lengths in the two strings.

I am afraid that I hardly have the right to say what the potential energy should be on Kelvin's theory, in respect to a point on which he has not expressed himself. It is however natural, if we imagine a sea of jelly to be surging to and fro among fixed molecules, to suppose that the jelly is thereby distorted. There are indeed two exceptions; 1st. If we suppose that the molecules do not affect the motion of the jelly, although they may have made it denser (by their attraction); 2nd. If we suppose the molecules to move with the jelly, taking the same displacements. In either of these cases (both apparently improbable) the momentum of the jelly would have the same ratio to the velocity for all directions about a point. This is not the case in Kelvin's theory, in which the effective inertia is aeolotropic (*Phil. Mag.* XXVI, 501). If then the jelly is distorted in its motion by the molecules among which it flows, the potential energy should not vanish for $l = \infty$, at least apparently not. (It is hard to speak with absolute certainty in regard to so peculiar a jelly, which enjoys moreover, I suppose, the privilege of further hypotheses in its behalf as they may become necessary.) I think we may say that it is reasonable to admit such a term in the formula on *a priori* evidence, its magnitude (whether zero or otherwise) to be determined *a posteriori*. I am not claiming the existence of an absolutely constant term, but one which does not vanish for $l = \infty$. I should expect, however, that it would be tolerably constant in most cases. As a matter of fact, it seems to me that we can measure it, as compared with the other term, by means of the dispersion of light, as I have in-

dicated in the passage in question. As thus measured, it does not appear at all constant (for constant amplitude) as I should expect it to be. This seems to me a very serious objection to the theory, although one which I should urge with great diffidence, as Lord Kelvin has not touched upon these points.

I thank you very much for your kind interest in my "Equilib. Het. Subst." I myself had come to the conclusion that the fault was that it was too *long*. I do not think that I had any sense of the value of time, of my own or others, when I wrote it.[3] Just now I am trying to get ready for publication something on thermodynamics from the *a priori* point of view, or rather on "Statistical Mechanics" of which the principle interest would be in its application to thermodynamics—in the line therefore of the work of Maxwell and Boltzmann. I do not know that I shall have anything particularly new in substance, but shall be contented if I can so choose my standpoint (as seems to me possible) as to get a simpler view of the subject.

I remain
Yours faithfully
J. Willard Gibbs

[1] *Life of Lord Rayleigh*, London, 1924. See pp. 172–173.

[2] As I interpret the statement of Lord Rayleigh in his *Scientific Papers*, vol. IV, p. 12, where the results of his experiment in question are reprinted, he finally concluded that the discrepancy between theory and experiment was not sufficient to require any explanation.

[3] This and the preceding two sentences are the three quoted in the Life of Lord Rayleigh. The reference to Statistical Mechanics which immediately follows is of particular interest in view of the publication about a decade later of the Elementary Principles in Statistical Mechanics.

STUDIES ON GEODESICS IN VIBRATIONS OF ELASTIC BEAMS

BY A. D. MICHAL

CALIFORNIA INSTITUTE OF TECHNOLOGY

Communicated December 12, 1944

The transverse vibrations of a homogeneous elastic beam that is simply supported (hinged) at both ends satisfy the fourth order partial differential equation

$$\frac{\partial^2 u(x, t)}{\partial t^2} + c^2 \frac{\partial^4 u(x, t)}{\partial x^4} = 0 \qquad (1)$$

and the boundary conditions

$$u(o, t) = \left.\frac{\partial^2 u(x, t)}{\partial x^2}\right]_{x=0} = 0, \; u(l, t) = \left.\frac{\partial^2 u(x, t)}{\partial x^2}\right]_{x=l} = 0, \qquad (2)$$

where the constant $c^2 = EI/\rho$, the ratio of the flexural rigidity EI and mass ρ per unit length of the beam, and l is the length of the beam.[1]

The kinetic and potential energies are given, respectively, by

$$T = \frac{\rho}{2}\int^{l}\left(\frac{\partial u\,(x,\,t)}{\partial t}\right)^2 dx,\ U = \frac{EI}{2}\int_0^{l}\left(\frac{\partial^2 u(x,\,t)}{\partial x^2}\right)^2 dx. \qquad (3)$$

It is well known that the differential equation (1) is the Euler-Lagrange equation resulting from *Hamilton's Principle*

$$\delta\int_{t_0}^{t_1}\Big(T - U\Big)dt = 0 \qquad (4)$$

when the usual restrictions are made on fourth order derivatives. There is a conservation of total energy so that $T + U = C$, a constant along a motion of the beam.

A one-parameter family of positions of the vibrating beam will be called a *dynamical path* or *dynamical trajectory*. A dynamical path need not be specified by the time parameter t. Let $u = u(x, t)$ be a dynamical path with the time parameter t and total energy C. Let us define a new parameter s along this dynamical path by

$$s = \int_0^t A(\lambda)d\lambda, \qquad (5)$$

where

$$A(\lambda) = 2C - EI\int_0^{l}\left(\frac{\partial^2 u(x,\,\lambda)}{\partial x^2}\right)^2 dx. \qquad (6)$$

In terms of this new parameter s, the conservation of total energy yields the condition

$$\rho\int_0^{l}\left(\frac{\partial \bar{u}(x,\,s)}{\partial s}\right)^2 dx = \frac{1}{\bar{A}(s)} \qquad (7)$$

on defining $\bar{u}(x, s) = u(x, t(s))$ and $\bar{A}(s) = A(t(s))$. Now condition (7) can be written in the following equivalent form:

$$\bar{A}^2(s)\frac{\rho}{2}\int_0^{l}\left(\frac{\partial \bar{u}(x,\,s)}{\partial s}\right)^2 dx + \frac{EI}{2}\int_0^{l}\left(\frac{\partial^2 \bar{u}(x,\,s)}{\partial x^2}\right)^2 dx = C. \qquad (8)$$

This then is the statement of the conservation of total energy when it is expressed in terms of the parameter s instead of the time t.

If the boundary conditions (2) are used in integrations by parts and if we employ the energy condition (7) or (8), we can establish the following theorem.

THEOREM 1. *If $u = u(x\ t)$ is a dynamical path with the time parameter t and total energy level C, then in terms of the parameter s defined in (5), it satisfies the integro-differential equation*

$$\frac{\partial^2 \bar{u}(x, s)}{\partial s^2} + \frac{1}{\bar{A}(s)}\left\{-2EI\left[\int_0^l \frac{\partial^4 \bar{u}(x_1, s)}{\partial x_1^4}\ \frac{\partial \bar{u}(x_1, s)}{\partial s}dx_1\right]\frac{\partial \bar{u}(x, s)}{\partial s} + EI\left[\int_0^l \left(\frac{\partial \bar{u}(x_1, s)}{\partial s}\right)^2 dx_1\right]\frac{\partial^4 \bar{u}(x, s)}{\partial x^4}\right\} = 0. \quad (9)$$

In other words, the partial differential equation (1) for the vibration states of a beam, simply supported at both ends and having *the same total energy level C*, becomes the integro-differential equation (9) when it is expressed in terms of the parameter s.

Let us now consider a converse problem to that of deriving (9). In fact, we shall start with the equation (9) and *assume* that $\bar{u}(x, s)$ is a solution of (9) with certain properties. If we define a parameter t by

$$t = \int_0^s \frac{d\lambda}{\bar{A}(\lambda)}, \quad (10)$$

where

$$\bar{A}(\lambda) = 2C - EI\int_0^l \left(\frac{\partial^2 \bar{u}(x, \lambda)}{\partial x^2}\right)^2 dx, \quad (11)$$

then one can prove the following result.[2]

THEOREM 2. *If $\bar{u}\ (x, s)$ is a solution of the integro-differential equation (9) such that $\bar{A}(s) > 0$, $\bar{u}(o, s) = \bar{u}(l, s) = 0$, $\left.\frac{\partial^2 \bar{u}(x, s)}{\partial x^2}\right]_{x=0} = \left.\frac{\partial^2 \bar{u}(x, s)}{\partial x^2}\right]_{x=l} = 0$, if s is a general parameter, C is a fixed positive number, and if $\bar{u}(x, s)$ satisfies the energy integral (7), or equivalently (8), then by the change (10) of parameter s to the time parameter t, the function $u(x, t)$ defined by $u(x, t) = \bar{u}(x, s(t))$ will satisfy the partial differential equation (1), the boundary conditions (2) and will have a total energy level C—in other words $u = u(x, t)$ will be a dynamical path in the time parameter t and will have a total energy level C.*

To bring out some geometrical ideas connected with our subject, let us multiply both sides of the energy integral (7) by $\bar{A}(s)\left(\frac{ds}{d\lambda}\right)^2$, where λ is an arbitrary (suitably admissible) parameter. This gives

$$2\left[C - \frac{EI}{2}\int_0^l\left(\frac{\partial^2 \bar{u}(x, s(\lambda))}{\partial x^2}\right)^2 dx\right]\rho\int_0^l\left(\frac{\partial \bar{u}(x_1, s(\lambda))}{\partial \lambda}\right)^2 dx_1 = \left(\frac{ds}{d\lambda}\right)^2. \quad (12)$$

Hence, if we define $v(x, \lambda)$ by $v(x, \lambda) = \bar{u}(x, s(\lambda))$ we obtain

$$s = \int_{\lambda_0}^{\lambda} \left\{ 2\left[C - \frac{EI}{2}\int_0^l \left(\frac{\partial^2 v(x, \lambda_1)}{\partial x^2}\right)^2 dx \right] \rho \int_0^l \left(\frac{\partial v(x_1, \lambda_1)}{\partial \lambda_1}\right)^2 dx_1 \right\}^{1/2} d\lambda_1. \tag{13}$$

The parameter s defined by (13) is, however, the arc length of a curve $v = v(x, \lambda_1)$, parameterized with parameter λ_1, $(\lambda_0 \leq \lambda_1 \leq \lambda)$, in an *infinitely dimensional "Riemannian" geometry* with function coordinates $v(x)$ and element of arc length squared given by

$$ds^2 = 2\left[C - \frac{EI}{2}\int_0^l \left(\frac{d^2 v(x)}{dx^2}\right)^2 dx \right] \rho \int_0^l \left(\delta v(x_1)\right)^2 dx_1, \tag{14}$$

where $\delta v(x)$ as well as $v(x)$ are arbitrary continuous functions with fourth order continuous derivatives such that $v(x)$ and $\delta v(x)$ as well as their second derivatives vanish at both ends of the beam, $x = 0$ and $x = l$. It is important to notice here that *the boundary conditions for a beam hinged at both ends play an important rôle in the determination of the function coordinate space of the geometry.*[3] We shall assume that $v(x)$ satisfies the inequality

$$C > \frac{EI}{2}\int_0^l \left(\frac{d^2 v(x)}{dx^2}\right)^2 dx,$$

a property that guarantees the positive definiteness of the quadratic functional differential form (14).

By definition, the curves $v = v(x, \lambda)$ that make the length functional (13) have a stationary value are the *geodesics* of our infinitely dimensional "Riemannian" space. The generalized Euler-Lagrange equation for this generalized calculus of variations problem can be shown to be the *integro-differential equation* (9) provided that the general parameter λ_1, is taken to be the arc length parameter s and provided $\bar{u}(x, s)$ is replaced throughout by $v(x, s)$.

If λ is the time t, $v(x, t)$ is a solution of the beam problem hinged at both ends and having total energy C, and if $T(t)$ is the *corresponding* kinetic energy at any time t, then it can be shown readily that the arc length s of the dynamical path $v = v(x, t)$ is given by

$$s = \int_0^t 2T(t_1)dt_1.$$

This result provides a physical interpretation for the arc length of a geodesic in our infinite dimensional "Riemannian" space. From another point of view, it can be regarded as a *generalization of Hamilton's characteristic function*[4] Hamilton introduced his characteristic function (action integral) in his studies on optics and then later in his studies on dynamical systems with a finite degree of freedom and with an energy integral.

It is clear from our previous discussion that the energy integral (8) for a solution of the integro-differential equation (9) is equivalent to the demand that the parameter s be an arc length parameter of the infinite dimensional "Riemannian" space (14). With the aid of this result, we can summarize Theorem 1 and Theorem 2 briefly in the language of vibrations as follows:

THEOREM 3. *The vibration states (dynamical path) with total energy level C of an elastic beam with both ends hinged can be represented by a geodesic in the infinite dimensional "Riemannian" space whose element of arc length squared is given by (14), and conversely.*

It can be shown that the infinite dimensional "Riemannian" space (14) is *not* of constant "Riemannian" curvature.[5] Hence the *vibrations of an elastic beam hinged at both ends with constant total energy furnish a physical model for the geodesics of a special type of infinite dimensional "Riemannian" space with variable "Riemannian" curvature.*

The harmonic vibrations of the simply supported beam are given by

$$u_n(x, t) = D_n \sin \frac{n\pi x}{l} \sin w_n t \ (n = 1, 2, \ldots, \ldots),$$

where the angular frequency

$$w_n = \frac{n^2\pi^2}{l^2}\sqrt{\frac{EI}{\rho}}.$$

A simple calculation shows that for $C > 0$ and for each integral value of n, the harmonic vibration

$$u_n(x, t) = \frac{2}{n^2\pi^2}\sqrt{\frac{l^3 C}{EI}} \sin \frac{n\pi x}{l} \sin\left(\frac{n^2\pi^2}{l^2}\sqrt{\frac{EI}{\rho}}t\right)$$

is a closed geodesic in the infinite dimensional "Riemannian" space with element of arc length ds given by (14). It can also be shown that this infinite system of closed geodesics $u_1(x, t), u_2(x, t), \ldots, \ldots$ forms a mutually orthogonal system of curves at the origin $v(x) = 0$ of the space (14) and that the tangent vectors $\frac{\partial u_n(x, t)}{\partial t}$ to these geodesics at the origin have the same magnitude given by twice the total energy C of each of these harmonic vibrations.

With the same general methods and obvious occasional changes in detail, one can treat the vibrations of cantilever beams, the vibrations of other beams, and many other problems in the mechanics of continuous media.[6]

[1] See Kármán and Biot, *Mathematical Methods in Engineering*, pp. 267–273, 283–290. See also Timoshenko, S., *Vibration Problems in Engineering*, p. 221.

[2] The method of proof and details are similar to those given by the author in another connection. See Michal, A. D., *The Vibrations of Elastic Strings as Studies in Geodesics*.

[3] The boundary conditions of functional equations play a similar geometrical rôle in many other connections. In the applications, this is not only the case in all sorts of vibration problems, but also in many other physical phenomena.

[4] See Whittaker, E. T., *Analytical Dynamics* (Cambridge, 1927), pp. 289 and 317. The reader who is acquainted with the Maupertuis-Jacobi "Least" Action Principle in finite degrees of freedom will notice that we have incidentally given a generalization of this principle for an infinite degree of freedom problem.

[5] The general tensor calculus and the geometric concepts associated with it can be improvised by anyone who is acquainted with the author's general "Riemannian" geometries as studied in Michal, A. D., "General Differential Geometries and Related Topics," *Bull. Amer. Math. Soc.*, **45** 529–563 (1939), especially pp. 551–559. References to the author's earlier work are also given in this paper.

[6] While giving an account of this paper in a lecture to my Seminar on Applied Mathematics, Bateman kindly called my attention to an application of Haar's work on the calculus of variations to Hamilton's Principle (4). Assuming only the existence of continuous second order partial derivatives, Hamilton's Principle leads to the *two* partial differential equations

$$\frac{\partial u(x, t)}{\partial t} = c^2 \frac{\partial^2 w(x, t)}{\partial x^2}, \quad \frac{\partial^2 u(x, t)}{\partial x^2} = -\frac{\partial w(x, t)}{\partial t} \tag{a}$$

in terms of an auxiliary function $w(x, t)$. If continuous fourth order derivatives are assumed to exist, the elimination of $w(x, t)$ in (a) leads to the classical equation (1) for beam vibrations. In terms of the parameter s, equations (a) become the *two integro-differential equations*

$$\frac{\partial \bar{u}(x, s)}{\partial s} = \frac{c^2}{\bar{A}(s)} \frac{\partial^2 \bar{w}(x, s)}{\partial x^2}, \quad \frac{\partial \bar{w}(x, s)}{\partial s} = -\frac{1}{\bar{A}(s)} \frac{\partial^2 \bar{u}(x, s)}{\partial x^2} \tag{b}$$

If continuous fourth order derivatives are assumed to exist, the elimination of $\bar{w}(x, s)$ in (b) leads to the fundamental integro-differential equation (9).

Similar remarks can be made in other problems such as the vibrations of an elastic string.

AN EXTENSION OF LIE'S THEOREM ON ISOTHERMAL FAMILIES

BY EDWARD KASNER AND JOHN DE CICCO

DEPARTMENTS OF MATHEMATICS

COLUMBIA UNIVERSITY AND ILLINOIS INSTITUTE OF TECHNOLOGY

Communicated December 7, 1944

1. Conformal Representation.—Let a surface Σ be represented conformally upon a plane π with cartesian coordinates (x, y). The linear-element of Σ is then of the form

$$ds^2 = E(x, y)(dx^2 + dy^2), \tag{1}$$

where $E(x, y) > 0$. The parametric curves $x =$ const. and $y =$ const. form an isothermal net on Σ.

It is well known that $h(x, y) = c$, where h is a harmonic function of (x, y), that is, h satisfies the Laplace equation, $h_{xx} + h_{yy} = 0$, defines an isothermal family of curves on Σ. The constant c is called the isothermal parameter.

The converse of the preceding statement is not valid. That is, if $g(x, y) =$ const. defines an isothermal family of curves on Σ with linear-element of the form (1), then it does not follow necessarily that g is a harmonic function, but it must be a function of a harmonic function. The exact statement which is due to Lie is as follows. The family of curves $g(x, y) =$ const. represents an isothermal system on the surface Σ with linear-element of the form (1), if and only if g satisfies the partial differential equation of third order

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) \arctan \frac{g_y}{g_x} = 0. \tag{2}$$

This means geometrically that the angle θ between the ∞^1 curves $g(x, y) =$ const. and the isothermal family $x =$ const. (or $y =$ const.) is a harmonic function of (x, y).

2. Statement of Our Problem.—We propose to give the necessary and sufficient condition that $g(x, y) =$ const., represent an isothermal family upon a surface Σ when (x, y) are general curvilinear coordinates on Σ. That is, when the linear-element of Σ is of the form

$$ds^2 = E(x, y)dx^2 + 2F(x, y)dxdy + G(x, y)dy^2, \tag{3}$$

where $H^2 = EG - F^2 > 0$, we shall determine the necessary and sufficient condition that $g(x, y) =$ const. represents an isothermal system of curves on Σ. This leads to a wide extension of the theorem of Lie, which is stated above.

The preceding condition is simpler when the parametric curves form an orthogonal net, that is, when $F = 0$. Of course, the simplest condition is (2) when the parametric curves form an isothermal net and (x, y) are isothermal coordinates.

As an application of our preceding work, we shall obtain the necessary and sufficient condition that the ∞^1 curves $g(x, y) =$ const. shall represent an isothermal family upon the Monge surface $\Sigma : z = f(x, y)$.

Finally the condition is found that the level curves $z =$ const. of the surface $\Sigma : z = f(x, y)$ be an isothermal family. This is applied to the mapping upon a plane π of the loxodromes (relative to the level curves) of the surface Σ, showing that they can be represented by straight lines for a sphere (Mercator) and a spheroid (Lambert), (and for any minimal surface and also for any surface of revolution with axis perpendicular to the xy-plane), but not for an ellipsoid of three unequal axes. Here use is made of a theorem of Kasner, which states that the complete system of ∞^2 isogonal trajectories of a given family of curves is linear if and only if the given family is isothermal.

3. *The Condition When the Minimal Lines Are Given in the Finite Form.*—Let (u, v) denote the minimal coordinates of any point on the surface Σ. Then $u = x + iy$, $v = x - iy$ where (x, y) are the isothermal coordinates defined in (1).

If a general point transformation is applied to the plane π upon which the surface Σ is conformally represented by means of the equation (1), it is found that u and v are given by the general expressions

$$u = \phi(x, y), \qquad v = \psi(x, y), \tag{4}$$

where ϕ and ψ are conjugate complex functions of the real variables (x,y). Of course, (x, y) are now general curvilinear coordinates of any point on Σ since the linear-element of Σ is of the form (3). The finite forms of the equations of the minimal lines are $\phi(x, y) =$ const. and $\psi(x, y) =$ const.

We seek the condition for an isothermal family in the general curvilinear coordinates (x, y). According to Lie's theorem, any isothermal family is defined in minimal coordinates (u, v) by a differential equation of the form

$$\log \frac{dv}{du} = \lambda(u) + \mu(v). \tag{5}$$

By substituting (4) into this equation, we find that the ∞^1 curves defined by the differential equation of first order $dy/dx = p = p(x, y)$ form an isothermal family if and only if the function p of (x, y) satisfies an equation of the form

$$\log \frac{\psi_x + p\psi_y}{\phi_x + p\phi_y} = \lambda(\phi) + \mu(\psi). \tag{6}$$

The functions λ of ϕ and μ of ψ must be eliminated by partial differentiation. Firstly upon applying the operation $\phi_y\partial/\partial x - \phi_x\partial/\partial y$ and simplifying, and secondly the operation $\psi_y\partial/\partial x - \psi_x\partial/\partial y$ to the above equation, we obtain the expression

$$\left(\psi_y \frac{\partial}{\partial x} - \psi_x \frac{\partial}{\partial y}\right)\left[\frac{(\phi_y(\partial/\partial x) - \phi_x(\partial/\partial y))\log \dfrac{\psi_x + p\psi_y}{\phi_x + p\phi_y}}{\phi_x\psi_y - \phi_y\psi_x}\right] = 0. \quad (7)$$

This condition is the necessary and sufficient condition that the curves defined by the differential equation $dy/dx = p(x, y)$ form an isothermal family on the surface Σ whose minimal curves are given in the finite form by $\phi(x, y) =$ const. and $\psi(x, y) =$ const.

4. *The Condition (7) in Terms of the Differential Equations of the Minimal Curves.*—For this purpose, let us define $\alpha(x, y)$ and $\beta(x, y)$ by the equations

$$\phi_x = -\alpha\phi_y, \quad \psi_x = -\beta\psi_y,$$

so that the differential equations of the minimal curves are $dy/dx = \alpha$ and $dy/dx = \beta$.

Upon substituting these into (7) and simplifying, we obtain the expression ultimately

$$\left[\frac{\partial^2}{\partial x^2} + (\alpha + \beta)\frac{\partial^2}{\partial x \partial y} + \alpha\beta\frac{\partial^2}{\partial y^2}\right]\log\frac{p - \beta}{p - \alpha} + \left(\frac{\beta_x + \alpha\beta_y}{\alpha - \beta}\right)\times$$
$$\left(\frac{\partial}{\partial x} + \alpha\frac{\partial}{\partial y}\right)\log\frac{p - \beta}{p - \alpha} - \left(\frac{\alpha_x + \beta\alpha_y}{\alpha - \beta}\right)\left(\frac{\partial}{\partial x} + \beta\frac{\partial}{\partial y}\right)\log\frac{p - \beta}{p - \alpha} +$$
$$(\beta\alpha_{yy} - \alpha\beta_{yy} + \alpha_{xy} - \beta_{xy}) + \left(\frac{\alpha_y - \beta_y}{\alpha - \beta}\right)(\alpha\beta_y - \beta\alpha_y + \beta_x - \alpha_x) = 0. \quad (8)$$

This is the necessary and sufficient condition that the family of curves defined by the differential equation $dy/dx = p = p(x, y)$ is an isothermal family when the minimal curves of the surface Σ are given by the differential equations $dy/dx = \alpha(x, y)$ and $dy/dx = \beta(x, y)$.

5. *The Condition (8) in Terms of the General Form (3) of the Linear-Element of the Surface Σ.*—If (x, y) are curvilinear coordinates on the surface Σ such that the linear-element of Σ is given by the general equation (3), it follows that α and β must be given by

$$\alpha = \frac{-1}{G}(F - iH), \quad \beta = -\frac{1}{G}(F + iH). \quad (9)$$

Let θ denote the expression

$$\theta = \text{arc cot } \frac{1}{H} (pG + F). \tag{10}$$

This is actually the angle θ between any curve of the family $dy/dx = p(x, y)$ and the parametric curves x = const.

Substituting the values of α and β as given by equations (9), we find that the equation (8) assumes the form

$$H\left[G\frac{\partial^2\theta}{\partial x^2} - 2F\frac{\partial^2\theta}{\partial x\partial y} + E\frac{\partial^2\theta}{\partial y^2}\right] + [HG_x - GH_x + FH_y - HF_y]\frac{\partial\theta}{\partial x} +$$
$$[FH_x - HF_x + \frac{1}{G}\{-EHG_y + 2FHF_y + (EG - 2F^2)H_y\}]\frac{\partial\theta}{\partial y} +$$
$$\frac{H}{G}[HF_{yy} - FH_{yy} + GH_{xy} - HG_{xy}] + \frac{1}{G^2}(GH_y + HG_y)(FH_y - HF_y) +$$
$$\frac{1}{G^2}(H^2G_xG_y - G^2H_xH_y) = 0. \tag{11}$$

This is our extension of Lie's theorem on isothermal familes. If it is desirable to write the above equation in a form which does not contain partial derivatives of H, we find the form

$$2H^2\left[G\frac{\partial^2\theta}{\partial x^2} - 2F\frac{\partial^2\theta}{\partial x\partial y} + E\frac{\partial^2\theta}{\partial y^2}\right] + [(EG-2F^2)G_x - G^2E_x + 2FGF_x +$$
$$EFG_y + GFE_y - 2EGF_y]\frac{\partial\theta}{\partial x} + [EFG_x + GFE_x - 2EGF_x - E^2G_y +$$
$$(EG - 2F^2)E_y + 2EFF_y]\frac{\partial\theta}{\partial y} + H[-FE_{yy} - \frac{EF}{G}G_{yy} + 2EF_{yy} + GE_{xy} -$$
$$\frac{1}{G}(EG - 2F^2)G_{xy} - 2FF_{xy}]$$
$$+\frac{1}{H}\begin{bmatrix} GFE_y^2 + \frac{EF}{G^2}(2EG - F^2)G_y^2 + 4EFF_y^2 + \frac{F}{G}(EG+F^2)\times \\ E_yG_y - (EG+3F^2)E_yF_y - \frac{E}{G}(3EG + F^2)G_yF_y - G^2E_xE_y - \\ F^2E_xG_y + 2FGE_xF_y + \frac{1}{G^2}(2H^4 - E^2G^2)G_xG_y - F^2E_yG_x + \\ 2EFG_xF_y + 2EFG_yF_x + 2GFE_yF_x - 2(EG + F^2)F_xF_y \end{bmatrix}$$
$$= 0. \tag{12}$$

6. The Condition (12) When the Parametric Curves Are Orthogonal.—The parametric curves are orthogonal if and only if $F = 0$. In that event the condition (12) becomes

$$2EG\left(G\,\frac{\partial^2\theta}{\partial x^2} + E\,\frac{\partial^2\theta}{\partial y^2}\right) + G(EG_x - GE_x)\,\frac{\partial\theta}{\partial x} + E(GE_y - EG_y)\,\frac{\partial\theta}{\partial y} +$$

$$(EG)^{1/2}(GE_{xy} - EG_{xy}) + (EG)^{-1/2}(E^2G_xG_y - G^2E_xE_y) = 0, \quad (13)$$

where $\theta = \text{arc cot}\, p(G/E)^{1/2}$.

This is the necessary and sufficient condition that the ∞^1 curves defined by the differential equation $dy/dx = p(x, y)$ be an isothermal family when the parametric curves on the surface Σ form an orthogonal net.

Of course, if the parametric curves form an isothermal net and if x and y are isothermal parameters, then (13) reduces to Lie's theorem stating that θ is a harmonic function.

7. The Condition (12) When the Surface Σ Is Given by the Monge Equation $z = f(x, y)$.—In this case $E = 1 + f_x^2$, $F = f_xf_y$, $G = 1 + f_y^2$. Substituting these values into (12), we obtain the expression

$$(1 + f_x^2 + f_y^2)\left[(1 + f_y^2)\,\frac{\partial^2\theta}{\partial x^2} - 2f_xf_y\,\frac{\partial^2\theta}{\partial x\partial y} + (1 + f_x^2)\,\frac{\partial^2\theta}{\partial y^2}\right] -$$

$$\left[(1 + f_y^2)f_{xx} - 2f_xf_yf_{xy} + (1 + f_x^2)f_{yy}\right]\left[f_x\,\frac{\partial\theta}{\partial x} + f_y\,\frac{\partial\theta}{\partial y}\right] + \frac{f_x(1+f_x^2+f_y^2)^{1/2}}{(1+f_y^2)} \times$$

$$[(1 + f_y^2)f_{xxy} - 2f_xf_uf_{xyy} + (1 + f_x^2)f_{yyy}]$$

$$+ (1 + f_x^2 + f_y^2)^{-1/2}\left[\begin{array}{l} -4f_xf_yf_{xy}^2 - 2\dfrac{f_xf_y(1 + f_x^2)}{(1 + f_y^2)^2}(2 + f_x^2 + 2f_y^2) \times \\ \quad f_{yy}^2 + (1 - f_x^2 + f_y^2)f_{xx}f_{xy} - 2f_xf_yf_{xx}f_{yy} + \\ \quad \dfrac{1}{(1 + f_y^2)^2}\left\{\begin{array}{l} 1+2f_y^2-f_x^4+f_y^4+9f_x^2f_y^2+3f_x^4f_y^2 \\ + 9f_x^2f_y^4 \end{array}\right\} \\ \qquad\qquad \times f_{xy}f_{yy} \end{array}\right]$$

$$= 0, \quad (14)$$

where $\theta = \text{arc cot}\,[p(1 + f_y^2) + f_xf_y]/[1 + f_x^2 + f_y^2]^{1/2}$.

The preceding equation is the necessary and sufficient condition that the ∞^1 curves defined by $dy/dx = p(x, y)$ be an isothermal family on the surface Σ which is given by the Monge equation $z = f(x, y)$.

8. The Condition That the Level Curves z = Const. Form an Isothermal Family.—As an application, we find that the condition that the level curves z = const. of the surface Σ defined by the Monge equation $z = f(x, y)$ be

an isothermal family is given by the equation (14) where $\theta = -\text{arc tan } f_y(1 + f_x^2 + f_y^2)^{1/2}/f_x$. Upon simplifying this, we find that the required third order condition is

$$\begin{aligned}(1 + f_x^2 + f_y^2)(f_x^2 + f_y^2)[f_y(1 + f_y^2)f_{xxx} - f_x(1 + 3f_y^2)f_{xxy} + f_y(1 + 3f_x^2)f_{xyy} \\ - f_x(1 + f_x^2)f_{yyy}] - 2(1 + 2f_x^2 + 2f_y^2)[(1 + f_y^2)f_{xx} - 2f_xf_yf_{xy} + (1 + f_x^2)f_{yy}] \\ \times [f_xf_yf_{xx} - (f_x^2 - f_y^2)f_{xy} - f_xf_yf_{yy}] = 0. \qquad (15)\end{aligned}$$

A first integral of this equation is

$$\frac{[(1 + f_y^2)f_{xx} - 2f_xf_yf_{xy} + (1 + f_x^2)f_{yy}]}{(f_x^2 + f_y^2)(1 + f_x^2 + f_y^2)} = \phi(f). \qquad (16)$$

Special classes of solutions are surfaces of revolution with axes perpendicular to the xy-plane, and cylinders with elements parallel to the xy-plane. All minimal surfaces are also solutions as may be verified by (16) since for a minimal surface the numerator of the fraction vanishes. We have proved that there are no surfaces such that the isothermal system of level curves on Σ may be represented on the xy-plane by similar ellipses or hyperbolas with the same axes, or congruent parabolas with the same axis. Hence the only quadric surfaces which belong to our class are those of revolution and the cylinders.

This may be applied to the mapping upon a plane π of the loxodromes (isogonals of the level curves) of the surface Σ. By a theorem of Kasner which states that the complete system of ∞^2 isogonal trajectories of a given family is linear (in the analytic sense) if and only if the given family is isothermal, it can be shown that the loxodromes may be represented by straight lines in the plane π for a sphere (Mercator) and a spheroid (Lambert), but not for an ellipsoid of three unequal axes. Also the loxodromes can be represented by straight lines in π for any minimal surface (in any orientation) and for any surface of revolution with axis perpendicular to π.

In conclusion, we may state that the problem of this section is equivalent to the determination of the class of surfaces Σ which can be projected orthogonally upon a plane π such that the unique Tissot net (the level curves together with their orthogonal trajectories) is isothermal. This suggests for consideration the more general problem of determining the class of surfaces Σ which can be pictured by a given non-conformal transformation T upon a given surface Σ_0 such that the unique orthogonal Tissot net determined by T on Σ or Σ_0 is isothermal. A variation of this problem is to have Σ and Σ_0 given, and then to determine all transformations T with the above property.

Other generalizations and applications of our fundamental formulas will be discussed elsewhere. We show that the only surfaces which are

intersected by every set of parallel planes in an isothermal family, besides the obvious cases of spheres and and planes, are the *minimal surfaces.*

This paper was presented before the American Mathematical Society, in April, 1944.

[1] Kasner, "Lineal Element Transformations Which Preserve the Isothermal Character," *Proc. Nat. Acad. Sci.*, **27**, 406–409 (1941).

[2] De Cicco, "Lineal Element Transformations Which Preserve the Dual-Isothermal Character," *Ibid.*, **27**, 409–412 (1941).

[3] Kasner, "Transformation Theory of Isothermal Families and Certain Related Trajectories," *Revista de Matematicas de Tucuman*, **2**, 17–24 (1941).

[4] Kasner and De Cicco, "Generalized Transformation Theory of Isothermal and Dual Families," *Proc. Nat. Acad. Sci.*, **28**, 52–55 (1942).

[5] Kasner and De Cicco, "Transformation Theory of Isogonal Trajectories of Isothermal Families," *Ibid.*, **28**, 328–333 (1942).

[6] Kasner and De Cicco, "An Extensive Class of Transformations of Isothermal Families," *Revista de Matematicas de Tucuman*, **3**, 271–282 (1942).

[7] De Cicco, "New Proofs of the Theorems of Beltrani and Kasner on Linear Families," *Bull. Amer. Math. Soc.*, **49**, 407–412 (1943).

[8] Kasner, "A Characteristic Property of Isothermal Systems," *Math. Ann.*, **59**, 252–354 (1904). *Geometry of isothermal systems*, Volume in honor of Rey Pastor, Instituto de Matematica, vol. 5, Rosaria, 1943.

NEW TYPES OF RELATIONS IN FINITE FIELD THEORY

BY H. S. VANDIVER

DEPARTMENT OF PURE MATHEMATICS, UNIVERSITY OF TEXAS

Communicated November 27, 1944

In two other recent papers[1] we developed methods which led to several new results in finite field theory, with particular application to ordinary congruences involving rational integers. In the present article we pursue these methods much further, obtaining various new kinds of relations.

The results just referred to gave criteria involving binomial coefficients for the number of roots of an equation in a finite field. To reduce these expressions to simpler forms in order to give more convenient criteria as to the number of roots, it seems necessary to go into considerations involving binomial coefficients which have not been heretofore studied, and so far we can only use these expressions to supply this information in comparatively few cases, some of which will be discussed elsewhere. At present we shall look at the situation from another angle. It is possible to apply the criteria to certain equations where we know in advance the number of roots or some properties of them. When this is done it turns out to be a fruitful method for finding relations of an entirely new type in number theory.

As one example of this, we obtain immediately from the statement of Theorem II of the first paper[1] and the remarks just preceding it, the result that the least residue, positive or zero, of

$$A = -c^2 \sum_{r=0}^{k} \sum_{k=1}^{m} \binom{kc}{rc} a^{rc} b^{mc-kc}, \tag{1}$$

modulo p, is $\leqq c$ for p any prime of the form $1 + cm$; m odd; a and b being any integers such that $(ab, p) = 1$. In this connection we note the remarks of H. J. S. Smith[2] in connection with a congruence due to Gauss, which states that if $p = 4n + 1 = h^2 + k^2$ is a prime then

$$h \equiv \pm \frac{(2n)!}{2(n!)^2} \qquad (\mathrm{mod}\ p). \tag{2}$$

Smith called this a remarkable relation since h may be determined directly by finding the least residue, in absolute value, of the right-hand member and this must be also $< \sqrt{p}$. The result concerning A in (1) is somewhat analogous.

1. We now note, as did Cipolla,[3] that we may introduce functions involving the roots of a certain congruence aside from merely the number of them.

For example, consider the expression,

$$N_r = \sum (x^r - x^r f(x)^{k(p^n - 1)}), \tag{3}$$

where the summation extends over all distinct elements x, $\neq 0$, of a finite field $F(p^n)$. For the values x_1 of x such that $f(x_1) = 0$, this expression reduces to x_1^r, and for an x not an x_1, it vanishes in the field.

Hence

$$N_r = \sum x_1^r$$

in $F(p^n)$ where the summation extends over all distinct roots x, of $f(x) = 0$. The relation (3) gives, if $r \not\equiv 0 \pmod{p^n - 1}$ and

$$(f(x))^{k(p^n-1)} = C_0 + C + C_2x^2 + \ldots;$$

$$N_r = C_t + C_{t+s} + C_{t+2s} \ldots, \tag{4}$$

where $t = p^n - 1 - r$; $s = p^n - 1$,

and

$$N_r + 1 = C_s + C_{2s} \ldots, \tag{5}$$

for $r \equiv 0 \pmod{s}$. For the latter case N_r equals, in the field, the number of roots of $f(x) = 0$. We shall find (4) convenient for further investigations.

From the point of view we are now employing, we shall find use also for the expression

$$\sum (x^r f(x)^d - x^r f(x)^{k(p^n-1)+d}), \tag{6}$$

where the summation extends over all distinct elements $\neq 0$ of $F(p^n)$ which contains (3), and $d > 0$. This function obviously reduces to zero in the field.

We now examine the relation

$$ax^m + 1 = 0; \quad p^n - 1 = mc; \tag{7}$$

and use the form (3) which gives in this case

$$N_r = \sum(x^{rm} - x^{rm}(ax^m + 1)^{k(p^n-1)}), \tag{8}$$

where the summation extends over all distinct values ρ such that $\rho^c = 1$ if we set ρ^{-1} in lieu of x^m.

Reducing (8) we obtain for $x^m = \rho^{-1}$,

$$\sum_\rho \rho^{-r} - \sum_\rho \sum_{s=0}^{k(p^n-1)} \binom{k(p^n-1)}{s} a^s \rho^{s-r}$$

or

$$-\sum_\rho \sum_{s=1}^{k(p^n-1)} \binom{k(p^n-1)}{s} a^s \rho^{s-r}.$$

Summation as to ρ gives

$$-c\sum_t \binom{k(p^n-1)}{t} a^t, \tag{9}$$

where t ranges over the integers in the set $1, 2, \ldots k(p^n - 1)$ such that $t \equiv r \pmod c$. We know from (7) that if $(-a)^c = 1$ then there is just one value x^m which satisfies the equation, and if $(-a)^c \neq 1$ there is none, that is, $N_r = (-1)^r a^r$ or 0 in (8), for $r > 0$. Assume also $r < c$. This gives

$$c\sum \binom{k(p^n-1)}{t} a^t = (-1)^{r+1} a^r \text{ or } 0, \tag{10}$$

according as $(-a)^c = 1$ or $(-a)^c \neq 1$, and if $t = 0$, then we must replace $(-1)^{r+1} a^r$ by -1 in this relation.

Now take the expression (6), and set $ax^m + 1$ for $f(x)$ and $(-r)$ for r Expansion gives, if $p^n - 1 > d > 0$,

$$c\sum_{t_1} \binom{d}{t_1} a^{t_1} - c\sum \binom{d + (p^n-1)k}{t} a^t = 0,$$

or

$$\sum_t \binom{d + (p^n-1)k}{t} a^t = \sum_{t_1} \binom{d}{t_1} a^{t_1}, \tag{11}$$

t_1 ranging over all the values in the set $0, 1, 2, \ldots, d$, such that $t_1 \equiv r \pmod c$, and t ranges over all the values in the set $0, 1, 2, \ldots, k(p^n - 1)$ such

that $t \equiv r \pmod{c}$. For $n = 1, a = 1, c = p^n - 1$ and $d \not\equiv 0 \pmod{p^n - 1}$ the relation (11) gives

$$\sum_t \binom{d + (p^n - 1)k}{t} \equiv \binom{d}{t} \pmod{p},$$

if $d < p^n - 1$, which is due to the writer,[4] and for $n = 1, t \equiv 0 \pmod{(p-1)}$ to Bachmann.[5]

Now set

$$\binom{v}{w} = 0,$$

for $w > v$, then relations (10) and (11) give the

THEOREM I. *If $p^n - 1 = mc$, p prime, $k > 0$, a is any element, $\neq 0$, of a finite field $F(p^n)$ then for $0 < r < c$,*

$$\sum_{s=0}^{\infty} c \binom{k(p^n - 1)}{r + cs} a^{cs} = (-1)^{r+1} a^r \text{ or } 0,$$

according as $(-a)^c = 1$, or $(-a)^c \neq 1$, also

$$\sum_{h=1}^{\infty} c \binom{k(p^n - 1)}{ch} a^{ch} = -1 \text{ or } 0,$$

with the same conditions on a.

For $d > 0$, $0 \leqq r < c$, then

$$\sum_{s=0}^{\infty} \binom{d + (p^n - 1)k}{r + cs} a^{cs} \equiv \sum_{s=0}^{\infty} \binom{d}{r + cs} a^{cs}.$$

2. To obtain certain other results we shall find it convenient here to introduce in the finite field of residue classes of a prime ideal (p) where p a primitive root of a prime l and use the algebraic field defined by a primitive lth root of unity designated by ζ. The ideal (p) is a prime ideal in said field. Set $p^{l-1} - 1 = lc$. Then if a is prime to p and also $a + 1 \not\equiv 0 \pmod{p}$, then $1 + a\zeta$ is also prime to p and

$$(1 + a\zeta)^{cr} \equiv \zeta^k \pmod{p}, \tag{14}$$

for some k, $0 \leqq k < l$.

Expand the left hand member and collect powers of ζ, then the result may be written

$$A_0 + A_1\zeta + \ldots + A_{l-1}\zeta^{l-1} \equiv 0 \pmod{p}. \tag{15}$$

We note that this congruence also holds if we set $\zeta^2, \zeta^3 \ldots, \zeta^{l-1}$ in lieu of ζ. Then we also note that

$$A_0 + A_1 + \ldots + A_{l-1} \equiv 0, \tag{16}$$

since c is divisible by $(p - 1)$ and $a + 1 \not\equiv 0 \pmod{p}$. The relations (15) and (16) after making the substitutions of the various powers of ζ already indicated give l congruences from which we may eliminate the A's since the determinant formed by the ζ's is an alternant which is prime to $p \neq l$. Hence

$$A_i \equiv 0; \qquad i = 0, 1, \ldots, l - 1.$$

Using the actual values of the A's we have

$$\binom{cr}{k} a^k + \binom{cr}{k + l} a^{k+l} + \ldots \equiv 1 \pmod{p},$$

and since $a \not\equiv 0 \pmod{p}$,

$$\binom{cr}{m} + \binom{cr}{km + l} a^l + \ldots \equiv 0 \pmod{p},$$

for $m \neq k$.

These give the

THEOREM II. *If p is a prime, and also a primitive root of a prime l, with $p^{l-1} - 1 = lc$, and r is any integer > 0, a is an integer with $(a(a + 1), p) = 1$, then for some k in the set $0, 1, \ldots, l - 1$, we have*

$$\sum_{s=0}^{\infty} \binom{cr}{k + sl} a^{k+sl} \equiv 1 \pmod{p},$$

and for any m in the set $0, 1, \ldots l - 1$ with $m \neq k$, we have

$$\sum_{s=0}^{\infty} \binom{cr}{m + sl} a^{sl} \equiv 0 \pmod{p}.$$

[1] These PROCEEDINGS, **30**, 362–367, 368–370 (1944).

[2] Smith, H. J. S., *Collected Math. Works*, v. 1, p. 269.

[3] Cipolla, *Periodico. di Mat.*, **22**, 36–41 (1907).

[4] *Ann. Math.*, II, **28**, 332 (1927).

[5] Bachmann, *Niedere Zahlentheorie*, II, **46** (1910).

FERMAT'S QUOTIENT AND RELATED ARITHMETIC FUNCTIONS

By H. S. Vandiver

Department of Pure Mathematics, University of Texas

Communicated November 27, 1944

If $(a, p) = 1$, p prime, the integer

$$q(a, p) = \frac{a^{p-1} - 1}{p} \tag{1}$$

has entered into many investigations in number theory. A number of questions concerning it are unsolved. For example, if a is fixed, is there a finite or infinite number of p's such that $q(a, p) = 0 \pmod{p}$, and if infinite is there a finite number of p's such that $a^{p-1} \equiv 1 \pmod{p^i}$ for i some given integer > 2?

Another somewhat related function is

$$W(p) = \frac{(p-1)! + 1}{p}, \tag{2}$$

known as Wilson's quotient. It differs from (1) in that it depends on p only. This function has a property[1] which connects it with the Bernoulli numbers, that is,

$$\frac{(p-1)! + 1}{p} \equiv b_{p-1} + \frac{p-1}{p} \pmod{p}, \tag{3}$$

where $(b + 1)^n = b_n$; $n > 1$, and after expansion we set $b^k = b_k$. Our object here is to introduce two other arithmetic functions (designated by G and M/p below) expressed as quotients, each of which is related to Fermat's quotient or the Bernoulli numbers.

1. We know that $b_k = 0$ for $k > 1$ and odd. Set $b_{2n} = (-1)^{n-1} B_n$ and further, write

$$B'_n = \frac{B_n}{n},$$

then consider the known relations

$$B'_{n+\mu} - B'_n \equiv 0 \pmod{p}; \quad \mu = (p-1)/2. \tag{4}$$

We then define the function

$$G(n, p) = \frac{B'_{n+\mu} - B'_n}{p}. \tag{5}$$

We have

$$G(n, p) \equiv G(n + \mu, p) \pmod{p},$$

so that there are only $p - 1$ incongruent G's modulo p. This follows from the relation[2]

$$B'_{n+y\mu} = yB'_{n+\mu} - (y - 1)B'_n.$$

We also infer from this that we get no essentially new functions by using $(B'_{n+k\mu} - B'_{n+s\mu})$ instead of $B'_{n+\mu} - B'_n$. We shall be interested here in the problem of transforming $G(n, p)$, modulo p, with $B_n \equiv 0 \pmod{p}$. There is no known case[2] where $G(n, p) \equiv 0 \pmod{p}$ under this restriction.

We have[3]

$$\frac{n^{2i} - 1}{2i} b_{2i} \equiv \sum_{a=1}^{p^2-1} y_a a^{2i-1} \pmod{p^2},$$

and

$$\frac{n^{2i+p-1} - 1}{2i + p - 1} b_{2i+p-1} \equiv \sum_{a=1}^{p^2-1} y_a a^{2i+p-2} \pmod{p^2},$$

and for $b_{2i} \equiv b_{2i+p-1} \equiv \pmod{p}$; $\quad i > 1$

$$\frac{n^{2i} - 1}{2i(2i - 1)} G(2i, p) \equiv \sum y_d d^{2i-1} q(d) \pmod{p}, \tag{6}$$

the summation extending over all values d in the set $1, 2, \ldots, p^2 - 1$ which are prime to p.

2. In another paper[4] the writer founded an arithmetical theory of the Bernoulli numbers on the congruence

$$S_n = 1^n + 2^n + \ldots + (p - 1)^n \equiv pb_n \pmod{p^2}$$

$n < p - 1$. Now S_n has the property that it is divisible by p for $n < p-1$. So also has the function

$$A = \sum_a a^n$$

where a ranges over all the positive integers, $a < (p - 1)$ such that $a^c \equiv 1 \pmod{p}$; also, $n \not\equiv 0 \pmod{c}$. However, all proofs of the above property of S_n seem to depend on additive properties of this function, whereas $\sum_a a$ is defined by means of multiplicative properties of certain integers and many of the additive properties of S_n and all apparently fail to carry over to A. Nevertheless, there is a relation between A/p and the Bernoulli numbers, and it is given in Theorem I below. This was a quite unexpected result. Consider the expression

$$\sum_{\rho} (1 - (a\rho + 1)^{p-1})^2 \equiv N \pmod{p^2}, \tag{7}$$

where ρ ranges over all the values such that $\rho^c \equiv 1 \pmod{p^2}$, where $p - 1 = mc$. Any term in this expression reduces to unity, modulo p^2, if ρ is such that

$$a\rho + 1 \equiv 0 \pmod{p}, \tag{8}$$

for $p > 2$. If this is not the case, however, then for such a value of the corresponding term (7) is divisible by p^2 Hence, N in (7) is the number of incongruent solutions ρ, modulo p, in (8). The number of such incongruent solutions is 1 or 0, according as $(-a)^c \equiv 1 \pmod{p}$ or $(-a)^c \not\equiv 1 \pmod{p}$. Expansion of (7) gives

$$N \equiv c - 2c\sum_{k=0}^{m} a^{kc}\binom{p-1}{kc} + c\sum_{h=0}^{2m} a^{hc}\binom{2(p-1)}{kc},$$

modulo p^2, and this may be written

$$N \equiv -2c\sum_{k=1}^{m} a^{kc}\binom{p-1}{kc} + c\sum_{k=1}^{m} a^{kc}\binom{2(p-1)}{kc} + c\sum_{t=1}^{m} a^{c(m+t)}\binom{2(p-1)}{(m+t)c}.$$

Multiply through by a^n, $n \not\equiv 0 \pmod{c}$, and using the value of N noted above, we have, after letting a range over the integers $1, 2, \ldots, p-1$ and adding, we obtain

$$\sum r^n \equiv -2c\sum_{k=1}^{m} S_{kc+n}\binom{p-1}{kc} + c\sum_{k=1}^{m} S_{kc+n}\binom{2(p-1)}{kc} + c\sum_{k=1}^{m} S_{cm+ck+n}\binom{2(p-1)}{cm+ck},$$

modulo p^2, where

$$S_i = 1^i + 2^i + \ldots + (p-1)^i,$$

and r ranges over the integers in the set $1, 2, \ldots, p-1$, such that $(-r)$ satisfies $x^c \equiv 1 \pmod{p}$. Using the formula, for i even,

$$S_i \equiv pb_i \pmod{p^2}, \tag{10}$$

and dividing (9) through by p, we obtain, modulo p, with c and n even,

$$\frac{\sum r^n}{p} \equiv -2c\sum_{k=1}^{m}\binom{p-1}{kc} b_{kc+n} + c\sum_{k=1}^{m}\binom{2(p-1)}{kc} b_{kc+n} + c\sum_{k=1}^{m} b_{ck+cm+n}\binom{2(p-1)}{cm+ck}. \tag{11}$$

It is known[5] that, modulo p,

$$\binom{\alpha p + \beta}{\alpha_1 p + \beta_1} \equiv \binom{\alpha}{\alpha_1}\binom{\beta}{\beta_1}$$

each of the integers α, β, α_1, β_1, being $\geqq 0$ and $< p$. Employing this, we note that, modulo p,

$$\binom{2(p-1)}{kc} \equiv \binom{p+p-2}{kc} \equiv \binom{p-2}{kc},$$

and

$$\binom{2(p-1)}{p-1+tc} \equiv \binom{p+p-2}{p+tc-1} \equiv \binom{p-2}{tc-1}.$$

Hence (11) reduces, modulo p, to

$$-2c\sum_{k=1}^{m} b_{kc+n}(-1)^{kc} + c\sum_{k=1}^{m} b_{kc+n}\binom{p-2}{kc} + c\sum_{t=1}^{m}\binom{p-2}{tc-1} b_{cm+ct+n}. \quad (12)$$

It is also known that

$$\binom{p-1}{r} \equiv (-1)^r \pmod{p}.$$

We also have

$$\binom{p-2}{kc} = \frac{(p-2)(p-3)\ldots\ldots(p-1-kc)}{2.\ 3.\ \ldots\ kc}.$$

Now

$$(p-2)(p-3)\ldots(p-(1+kc)) \equiv (-1)^{kc}\, 2, 3, \ldots (kc+1).$$

so that

$$\binom{p-2}{kc} \equiv (-1)^{kc}(kc+1) \pmod{p}.$$

Also

$$\binom{p-2}{tc-1} = \frac{(p-2)(p-3)\ldots\ldots(p-tc)}{2.\ 3\ \ldots\ (tc-1)}$$

$$\equiv (-1)^{tc-1}\, tc \pmod{p},$$

hence (12) becomes

$$-2c\sum_{k=1}^{m} b_{kc+n}(-1)^{kc} + c\sum_{k=1}^{m} b_{kc+n}(-1)^{kc}(kc+1) + c\sum_{t=1}^{m}\binom{p-2}{tc-1} b_{p-1+ct+n}. \quad (13)$$

Now use the relation

$$\frac{b_{s+p-1}}{s+p-1} \equiv \frac{b_s}{s} \pmod p$$

for $s \not\equiv 0 \pmod{p-1}$, we may then write, modulo p, if $M'(c, p, n)$ is $\sum_r r$ where r ranges over the integers h such that $(-h)^c \equiv 1 \pmod p$

$$\frac{M'(c, p, n)}{p} \equiv c\sum_{k=1}^{m}(-1)^{kc}(kc+1)\, b_{kc+n} + c(-1)^{kc-1}\sum_{k=1}^{m} kc\frac{kc+n-1}{kc+n}\, b_{kc+n} - 2c\sum_{k=1}^{m} b_{kc+n}(-1)^{kc},$$

and collecting the terms on the right we find

$$\frac{M'(c, p, n)}{p} \equiv cn\sum_{k=1}^{m}(-1)^{kc-1}\frac{b_{kc+n}}{kc+n}, \quad (14)$$

modulo p.

Now, since c and n are even,

$$\frac{M'(c, p, n)}{p} \equiv \frac{M(c, p, n)}{p} \pmod p,$$

where $M = \sum_a r^n$ where r ranges over the incongruent integers $< p$ which satisfy $x^c = 1$, hence we may simplify (14) and obtain

THEOREM I. *If p is an odd prime with $p - 1 = mc$, and $M(c, n, p)$ is defined by*

$$\frac{\sum r^n}{p},$$

where the summation extends over all integers r in the set $1, 2, \ldots, p-1$ such that $r^c \equiv 1 \pmod p$, with $n \not\equiv 0 \pmod c$, then if c and n are even,

$$M(c, n, p) \equiv -cn\sum_{k=0}^{m-1}\frac{b_{kc+n}}{kc+n} \pmod p. \quad (15)$$

For $c = p - 1$, then $m = 1$ and the result becomes the well-known relation (10).

It is possible to extend the Theorem I to the sum

$$\sum \frac{(ma + k)^n}{p},$$

where a ranges over the integers in the set $1, 2, \ldots, p - 1$ such that $a^c \equiv 1 \pmod{p}$, and so that the b's are replaced by numbers of the form $(mb + h)^{kc+n}$, but the proof will not be given here. Also it follows from Theorem I that

$$M(c, n, p)/n \equiv M(c, n + c, p)/(n + c) \pmod{p}.$$

[1] A proof is given by the writer, as a special case of an explicit expression for any Bernoulli number, in *Duke Math. Jour.*, **8**, 578 (1941), relation (26).

[2] Pollaczek, *Math. Zeit.*, **21**, 36 (1924).

[3] Vandiver, *Duke Math. Jour.*, **5**, 549 (1939).

[4] *Trans. Amer. Math. Soc.*, **51**, 510 (1942).

[5] Lucas, *Amer. Jour. Math.*, **1**, 229, 230 (1878).

NON-LINEAR INTEGRAL EQUATIONS OF THE HAMMERSTEIN TYPE

By C. L. Dolph

Department of Mathematics, Princeton University

Communicated November 14, 1944

Sufficient conditions for the existence and uniqueness of a solution to the integral equation:

$$\psi(x) = \int_a^b K(x, y) f\{y, \psi(y)\} dy \tag{1}$$

were investigated by Hammerstein[1] and his pupils. If $K(x, y)$ is a symmetric, positive-definite continuous kernel* in L^2 and $f(x, y)$ is continuous in x and y, their results may be classified into three closely related categories, according to the degree of non-linearity permitted $f(x, y)$

Hammerstein obtained the most general theorem by requiring that (x, y) satisfy the inequality:

$$\int^v f(x, y) dy \leq \mu_1 \frac{v^2}{2} + C_1; \quad C_1 > 0; \quad 0 < \mu_1 < \lambda_1 \tag{A}$$

for all v. Here λ_1 is the smallest characteristic value of $K(x, y)$ in the sense of linear integral equations. Under this condition Hammerstein established the existence of at least one function $\psi(x)$ satisfying (1). His method was essentially that of the Rayleigh-Ritz process, (A) being sufficient to guarantee the existence of a lower bound to the functional possessing (1) as its Euler-Lagrange equation.

Iglisch[2] imposed a stronger condition on $f(x, y)$; namely,

$$0 < \lim_{|y| \to \infty} \operatorname{Sup} \frac{f(x, y)}{y} < \mu_1 < \lambda_1. \tag{B}$$

Utilizing reasoning similar to that of the fixed-point method of Schauder-Leray, Iglisch proved the existence of a solution under (B).

Both Hammerstein and Iglisch also treated (1) under a still stronger condition on $f(x, y)$. In the event that $f(x, y)$ satisfied the inequality:

$$\left|\frac{f(x, y_2) - f(x, y_1)}{y_2 - y_1}\right| \leqslant \mu_1 < \lambda_1 \text{ for any } y_1, y_2, \tag{C}$$

both men were able to establish the existence and uniqueness of a solution to (1). Golomb[3] also applied the Picard Approximation Process to construct a solution under (C) even when $K(x, y)$ was unsymmetric and admitted an absolutely and uniformly convergent bilinear expansion in its characteristic functions.

In the event that $f(x, \psi)$ is a linear function in ψ, the conditions (A), (B) and (C) are all equivalent and the above results are classical. As is well known, in the linear theory it is not essential to limit the above inequalities to the smallest characteristic value. In fact all that is required is that the linear function f be bounded from above and below by consecutive characteristic values λ_n, λ_{n+1}. It is therefore natural to consider the non-linear integral equation (1) under the corresponding more general conditions: for all v

$$\mu_n \frac{v^2}{2} + Cn \leqslant \int_0^v f(x, y)\, dy \leqslant \mu_{n+1} \frac{v^2}{2} + C_{n+1} \tag{A$'$}$$

$$0 < \lambda_n < \mu_n < \mu_{n+1} < \lambda_{n+1}$$

$$\mu_n \leqslant \lim_{|y| \to \infty} \inf \frac{f(x, y)}{y} \leqslant \lim_{|y| \to \infty} \operatorname{Sup} \frac{f(x, y)}{y} \leqslant \mu_{n+1} \tag{B$'$}$$

$$\mu_n \leqslant \left|\frac{f(x, y_2) - f(x, y_1)}{y_2 - y_1}\right| \leqslant \mu_{n+1} \tag{C$'$}$$

for any y_2, y_1.

In the cases (B′) and (C′) the theorem corresponding to (B) and (C) and their methods of proof can be readily generalized. These proofs depend upon the existence of an *a priori* estimate of the norm of all possible solutions to (1) under the above conditions of $f(x, y)$. In order to establish an estimate, use was made of the fact that (B′), and hence *a fortiori* (C′), imply that $f(x, y)$ can be written as:

$$f(x, y) = \left(\frac{\mu_{n+1} + \mu_n}{2}\right)y + g(x, y) \tag{2}$$

where for sufficiently large N, $|y| > N$

$$|g(x, y)| < {}^1/_2\{\mu_{n+1} - \mu_n\}|y|. \tag{3}$$

This enables (1) to be written as one of the equations of the family defined by:

$$\psi_\tau(x) = \frac{(\mu_{n+1} + \mu_n)}{2} \int_a^b K(x, y)\psi_\tau(y)dy + \tau \int_a^b K(x, y)g[y, \psi_\tau(y)]dy. \tag{4}$$

That is, for $\tau = 1$, equation (4) reduces to (1); and for $\tau = 0$, equation (4) is an ordinary Fredholm equation of the first kind, which, since $\frac{\mu_{n+1} + \mu_n}{2}$ is distinct from all the characteristic values of $K(x, y)$, possesses only the solution $\psi_0(x) = 0$. Moreover, for any value of τ,

$$||\psi_\tau||^2 \leqslant \frac{A^2}{\delta^2 - \frac{\tau}{2}[\mu_{n+1} - \mu_n]^2} \leqslant \frac{A^2}{\delta^2 - \frac{[\mu_{n+1} - \mu_n]^2}{2}} \tag{5}$$

where

$$||g(\psi_\tau)|| \leqslant A \text{ when } ||\psi_\tau|| < N \tag{6}$$

and

$$\delta^2 = \operatorname{Min}\left\{\left[\lambda_{n+1} - \frac{(\mu_{n+1} + \mu_n)}{2}\right]^2, \left[\lambda_n - \frac{(\mu_{n+1} + \mu_n)}{2}\right]^2\right\}. \tag{7}$$

The fixed-point method of Schauder-Leray[4] can now be applied and the existence of a solution concluded under (B′) almost immediately. The estimate given by (5) is also entirely sufficient to prove the uniqueness of the solution under (C′).

In distinct contrast to this, no scheme yielded an *a priori* estimate under (A′). In fact, it is an easy matter to construct an example demonstrating that there is no topological reason for such an estimate to exist under these conditions. Thus the local methods of the calculus of variations have to be replaced by those in the large before further progress is possible. The generalization of (A) which was finally achieved is very incomplete, even though it does employ the methods in the large in a somewhat different way than they had hitherto been used. The method of proof does not in any way depend upon the existence of an *a priori* estimate, although, remarkably enough, once the existence of a solution has been established, it does permit an estimate of the norm of the solution to be made (*a posteriori*

estimation). Before stating these results and sketching the method of proof, it is convenient to state that equation (1) is equivalent to

$$\phi(x) = \int_a^b H(x, y) f\{y \int_a^b H(y, s) \psi(s)ds\}dy \tag{8}$$

where

$$K(x, y) = \int_a^b H(x, s) H(s, y)\, ds \tag{9}$$

in that if $\psi(x)$ is a solution of (1), then

$$\phi(x) = \int_a^b H(x, y) f[y, \psi(y)]dy \tag{10}$$

is a solution of (B), and conversely, if $\phi(x)$ is a solution of (8), then

$$\psi(x) = \int_a^b H(x, y)\, \phi(y)dy \tag{11}$$

is a solution of (1). Furthermore, equation (8) is the Euler-Lagrange equation of the functional

$$J(\phi) = (\phi, \phi) - 2G(H\phi) \tag{12}$$

where

$$G(H\phi) = \int_a^b x \int_0^{\int_a^b H(x, y)\phi(y)dy} f(x, v)dv \tag{13}$$

THEOREM: If $f(x, y)$ satisfies

$$\mu_n(\phi, \phi) - C_n \leqslant G(\phi) \leqslant \mu_{n+1}(\phi, \phi) + C_{n+1} \tag{A''}†$$

where μ_n, μ_{n+1} are less the nth and $n + 1$st characteristic values of $K(x, y)$ respectively; if

$$|f(x, y)| < A\, |y| + B \tag{14}$$

for some constants $A > 0$, $B > 0$; and if the functional $J(\phi)$ possesses only one maximum on any n-dimensional linear manifold in Hilbert space parallel to the one containing the origin and the first n characteristic functions of $K(x, y)$; then equation (1) has at least one solution.

Sketch of the Proof: Let Γ denote the class $\{M_n\}$ of n-dimensional manifold described above, M_n^0 the particular one containing the origin. Define

$$d(M_n) = \operatorname*{Sup}_{\phi \in M_n} J(\phi) \tag{15}$$

and

$$d = \inf_{M_n \in \Gamma} d(M_n). \tag{16}$$

Condition (A″) implies that the number d defined by (16) is finite. Each M_n^m is uniquely defined by its intersection with the manifold orthogonal

to M_n^0. This intersection will be denoted by ψ^α. For all manifolds such that

$$d(M_n^\alpha) < d(M_n^0) \tag{17}$$

the norm of the function ψ^α is governed by the inequality

$$||\psi^\alpha|| < G(M_n^0). \tag{18}$$

Thus, if a sequence of manifolds $\{M_n^\beta\}$ is chosen so that

$$d(M_n^\beta) \to d \tag{19}$$

the corresponding sequence $\{\psi^\beta\}$ will converge to a value, ψ^* since any bounded set of Hilbert space is sequentially compact in the usual weak topology [in which neighborhoods are defined in terms of an ϵ and any finite set of linear functions**].

The linear manifold M_n^* containing ψ^* and parallel to M_n^0 is a limit manifold for which

$$d(M_n^*) = d \tag{20}$$

and on which, by hypothesis, there exists a point ϕ_0 such that

$$J(\phi_0) = d. \tag{21}$$

Furthermore, unless

$$\text{grad } J(\phi_0) = 0 \tag{22}$$

it would be possible to displace M_n^* into a new manifold M_n^{**} on which

$$d(M_n^{**}) < d \tag{23}$$

contrary to the definition of d. One merely has to consider the displacement defined by

$$\phi_\tau = \phi_0 + \tau \text{ grad } J\ (\phi_0) \tag{24}$$

and to divide M_n^* into three concentric regions located around ϕ_0 as center. The first of these is defined as the region of ϕ_0 where grad $J(\phi)$ is uniformly bounded away from zero, the third as the region at infinity where

$$J(\phi) < \frac{d}{2} \tag{25}$$

and the second as the difference between the first and the third. The existence of the third region follows directly from the original hypotheses. The problem then reduces to a finite dimensional problem in n-variables and it follows by an examination of all possible cases that (24) would permit M_n^* to be displaced into M_n^{**}, satisfying (23). From this contradiction and (8) and (11), it follows that equation (1) has at least one solution. Moreover, since

$$||\psi^*|| < G(M_n^0) \tag{26}$$

and

$$J(\phi_0) = d, \ \phi_0 = \theta + \psi^* \tag{27}$$

an estimate of θ, and hence of ϕ_0, can be made with the aid of (A″).

In the event that $f(x, y)$ is linear and of the form

$$f(\phi) = \lambda\phi(x) + h(x) \tag{28}$$

this method of proof also yields an interesting geometrical interpretation of the relationship between the homogeneous and nonhomogeneous linear Fredholm problems which agrees with the one first given by Lusternik and Schnirelman[5] for the finite-dimensional analogue of the homogeneous case. It can be shown that the usual characteristic functions and values are given by considering linear cycles in the projective Hilbert space, which may be thought of as lying on a sphere at infinity with its diametrical opposite points identified and that the solutions to the non-homogeneous problem for values of λ between successive characteristic values may be obtained as above by considering n-dimensional linear manifolds passing through Hilbert space and parallel to M_n^0.

The complete and detailed proofs of the above results will be published in a forthcoming paper.

* The theorems are still true if the kernels are only "brauchbar unstetig."

† (A″) on equation (8) is easily seen to be the same as (A′) for equation (1).

** See, e.g., F. H. Murray, *Linear Transformations in Hilbert Space*, Princeton University Series.

[1] Hammerstein, A., *Acta Mathematica*, **54**, 117–176 (1930).

[2] Iglisch, R., *Math. Ann.*, **101**, 98–119 (1929).

[3] Golomb, M., *Math. Zeitschrift*, **39**, 45–75 (1935).

[4] Leray-Schauder, *Paris, Ecole normale supérieure Annales*, **51**, 45–78 (1934).

[5] Lusternik and Schnirelman, *Méthodes Topologiques dans les Problemes Variationnels, Actualités Scientifiques et Industrielles*, V. **188** (1934).

PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 31 February 15, 1945 Number 2

CONCERNING TANGENTS TO CONTINUA IN THE PLANE

By R. L. Moore

Department of Pure Mathematics, University of Texas

Communicated November 17, 1944

The line k is said[1] to be tangent to the non-degenerate continuum[2] M at the point P of M if every subset of M which has P as a limit point intersects every domain which contains $k-P$ and is the sum of the interiors of two vertical angles with vertex at P.

The straight line ray PB is said to be tangent to the non-degenerate continuum M at the point P of M if every subset of M which has P as a limit point intersects every domain which contains B and is the interior of an angle with vertex at P.

Theorem 1. *No continuum M contains an uncountable point set H such that for each point P of H there is a ray starting from P and tangent to M at P.*

Proof. Suppose there exists a continuum M containing such a point set H. If P is a point of H there exist an acute angle α with vertex at P and a domain D containing P such that $M \cdot D-P$ lies in the interior of α. It follows, by a theorem of Denjoy's, that H is countable.

Theorem 2. *If the non-degenerate continuum M has a tangent at each of its points then every subcontinuum of M is a continuous curve.*

Proof. Suppose N is a non-degenerate subcontinuum of M. Clearly every tangent to M at a point of N is also tangent to N at that point. Let K denote the set of all points P of N such that no ray starting from P is tangent to N at P. Every point of K is a local separating point of N. Hence, by Theorem 1 and a theorem of G. T. Whyburn's,[3] N is a continuous curve.

Theorem 3. *If the compact continuum M has a tangent at each of its points and K is the set of all emanation points of triods lying in M then $\bar{K}$ is totally disconnected.*

Proof. Suppose, on the contrary, that $\bar{K}$ contains a nondegenerate continuum and therefore an arc t. Since every point of t is a limit point of K, every one is a limit point of a subset H of K such that every point of H

is the emanation point of some simple triod lying wholly in t except for one arc that has only P in common with t. Suppose P_1 is a point of H. There exist three arcs P_1X_1, P_1X_2 and P_1X_3 lying in M such that two of them lie on t and the third one has only P_1 in common with t and no two of them have in common with each other any point except P_1. Let P_1A and P_1B denote the two rays which start from P_1 and lie on the straight line which is tangent to M at P_1. There exists a rectangle $CDEF$ of diameter less than 1 such that (1) its diagonals CE and DF intersect at P_1, (2) every point of $M-P_1$ that lies in its interior lies in the interior of one of the two triangles CP_1D and EP_1F, (3) X_1, X_2 and X_3 are in its exterior and (4) the angle CP_1D of the triangle CP_1D is acute. For each n less than 4 let Y_n denote the first point in the order from P_1 to X_n that the arc PX_n has in common with the perimeter of $CDEF$. Each of the points Y_1, Y_2 and Y_3 lies on one of the straight line intervals CD and EF. Hence there exist points C_1, D_1, W_1 and Z_1 such that (1) C_1 and D_1 are either C and D or E and F and W_1 and Z_1 are two of the points Y_1, Y_2 and Y_3 and (2) W_1 and Z_1 both lie on C_1D_1, (3) Z_1 belongs to t. Let P_1W_1 denote the interval, with end-points at P_1 and W_1, of that one of the arcs P_1X_1, P_1X_2 and P_1X_3 on which W_1 lies and let P_1Z_1 denote the interval, with end-points at P_1 and Z_1, of the one on which Z_1 lies. The arcs P_1W_1 and P_1Z_1 lie, except for their end-points, wholly in the interior of the triangle $C_1P_1D_1$.

Similarly, in view of the fact that P_1Z_1 is a subset of t, it is clear that there exist a triangle $C_2P_2D_2$ of diameter less than $1/2$ and arcs P_2Z_2 and P_2W_2 lying in M such that (1) if P is any point in the interior of the triangle $C_2P_2D_2$ the distance from P to P_1 is less than one half the distance from P to the line C_1D_1, (2) the angle $C_2P_2D_2$ of this triangle is less than one half of the angle $C_1P_1D_1$ of the triangle $C_1P_1D_1$, (3) the triangle $C_2P_2D_2$ lies wholly in the interior of the triangle $C_1P_1D_1$, (4) P_2 belongs to H and the arc P_2Z_2 is a subset of the arc P_1Z_1 and (5) W_2 and Z_2 lie on the interval C_2D_2 and each of the arcs P_2Z_2 and P_2W_2 lies, except for its end-points, wholly in the interior of the triangle $C_2P_2D_2$. This process may be continued. It follows that there exist an infinite sequence of triangles $C_1P_1D_1$, $C_2P_2D_2$, ... and two infinite sequences of arcs P_1W_1, P_2W_2, ... and P_1Z_1, P_2Z_2, ... such that $C_1P_1D_1$, P_1Z_1 and P_1W_1 are as described above and, if for each n, the angle $C_nP_nD_n$ of the triangle $C_nP_nD_n$ is called α_n and the interior of this triangle is called I_n then, for each n, (1) the diameter of I_n is less than $1/n$ and I_{n+1} is a subset of I_n, (2) the angle α_{n+1} is less than one-half of the angle α_n, (3) the distance from each point of I_{n+1} to the line C_nD_n is more than n times the distance from that point to P_n, (4) the arcs P_nZ_n and P_nW_n lie in M and, except for their end-points, in I_n and the points Z_n and W_n lie on the interval C_nD_n. The point sets I_1, I_2, I_3, ... have a point O in common. Suppose h and k are two straight lines through O and suppose there exist infinitely many positive integers n such that h

and k both intersect the interval C_nD_n. Then there exist two rays OL and OT lying on h and k, respectively, and such that, for infinitely many integers n, OL and OT both intersect C_nD_n and therefore the acute angle C_nOD_n is not less than the acute angle LOT. In view of conditions (2) and (3) this is clearly impossible. Hence there exists a line δ through O such that no other line through O intersects more than a finite number of the intervals of the sequence $C_1D_1, C_2D_2, \ldots$. Suppose x is a line through O distinct from q. There exists a number δ_x such that, for every n greater than δ_x, x contains no point of the side C_nD_n, and therefore intersects both of the sides P_nC_n and P_nD_n of the triangle $C_nP_nD_n$. It follows that, for each such n, x intersects both of the arcs P_nZ_n and P_nW_n in the interior of $C_nP_nD_n$ and thus contains two points of M (and therefore a point of M distinct from O) in the interior of that triangle. Thus every line through O except q contains a subset of M having O as a limit point. But this is impossible since there is a tangent to M at O.

THEOREM 4. *If the closure of the compact and countable point set K is totally disconnected then there exists a dendron M such that* (1) *K is the set of all junction points of M and* (2) *M is topologically equivalent to a dendron that has a tangent at every one of its points.*

Proof. Since $\bar{K}$ is totally disconnected and K is compact and countable there exists an arc α containing $\bar{K}$ and such that (1) every point of K is a limit point of some component of $\alpha-\bar{K}$ and (2) if K is nondegenerate the end-points of α belong to $\bar{K}$. There exists a reversibly continuous transformation T of the plane into itself such that $T(\alpha)$ is a straight line interval AB. Let $P_1, P_2, \ldots$ denote the points of $T(K)$ and for each n let K_n denote one component of $AB-T(\bar{K})$ of which P_n is a limit point. For each n, K_n is a straight line segment having P_n as one of its end-points. There exists a sequence of mutually exclusive circles $\beta_1, \beta_2, \beta_3, \ldots$ such that (1) for each n, β_n is tangent, at P_n, to the straight line that joins A and B, (2) the length of the segment K_n is greater than n times the diameter of β_n. For each n let γ_n denote an arc of β_n such that (1) one of its end-points is at P_n, (2) its length is less than one-fourth of the length of β_n, (3) its orthogonal projection onto AB is a subset of $P + K_n$. Let N denote the point set $AB + \gamma_1 + \gamma_2 + \ldots$. This point set is a dendron with a tangent at each of its points and K is the set of all junction points of the dendron of which N is the image under the transformation T.

THEOREM 5. *If the compact continuum M has a tangent at every one of its points and K is the set of all points P of M such that there is a straight line ray starting from P and tangent to M there then K is a countable inner limiting set and its closure is totally disconnected.*

Proof. Let K_1 denote the set of all end-points of M. The set K_1 is a subset of K. Let K_2 denote $K-K_1$. By Theorem 1, K_2 is countable. By a theorem of Menger's,[4] K_1 is an inner limiting set. Suppose K_2 is not.

Then it contains a point set K_3 which is dense in itself. Let P_1 denote some point of K_3. Let P_1O_1 denote the ray starting from P_1 which is tangent to M at that point. Let R_1 denote a circular domain with center at P_1 and diameter 1. There exists an isosceles triangle $C_1P_1D_1$ lying, together with its interior I_1, in R_1 and such that the point O_1 is within the acute angle C_1P_1-D_1 of this triangle and no point of M except P_1 is on either of the straight line intervals P_1C_1 and P_1D_1 but M contains two arcs P_1Z_1 and P_1W_1 having only P_1 in common and lying wholly in I_1 except for the point P and the points Z_1 and W_1 both of which lie on the interval C_1D_1. There exists a domain R_2 lying in R_1 and containing P_1 but no point of M that is not in the interior of the angle $C_1P_1D_1$. Since P_1 is a limit point of K_3, this domain contains a point P_2 of K_3 distinct from P_1 and at a distance from P_1 less than half its distance from the line C_1D_1. The point P_2 belongs to I_1. If P_2O_2 is the straight line ray which is tangent to M at P_2, and which starts at that point, there exists an isosceles triangle $C_2P_2D_2$, lying with its interior I_2 wholly in I_1 and within a circle with center at P_2 and diameter $^1/_2$, such that (1) O_2 lies within the angle $C_2P_2D_2$ of this triangle and no point of M except P_2 is on either of the straight line intervals P_2C_2 and P_2D_2 but M contains two arcs P_2Z_2 and P_2W_2 having only P_2 in common and lying wholly in I_2 except for P_2 and the points Z_2 and W_2 which lie on the interval C_2D_2, (2) the angle $C_2P_2D_2$ is less than one half the angle $C_1P_1D_1$ and (3) the distance from any point of I_2 to the line C_1D_1 is more than twice its distance from P_1. This process may be continued. It follows that there exist an infinite sequence of triangles $C_1P_1D_1$, $C_2P_2D_2$, ... and two infinite sequences of arcs P_1Z_1, P_2Z_2, ... and P_1W_1, P_2W_2, ... such that $C_1P_1D_1$ and P_1Z_1 and P_2Z_2 are as indicated above and such that if, for each n, I_n denotes the interior of the triangle $C_nP_nD_n$ and α_n denotes the angle $C_nP_nD_n$ of this triangle then, for each n, there hold true four conditions which may be worded precisely as are Conditions (1)–(4) concerning the I_n, α_n, P_nZ_n, P_nW_n, $C_nP_nD_n$, etc., with which the above proof of Theorem 3 is concerned. A contradiction follows as in that proof. Hence K_2 is an inner limiting set. Since both K_1 and K_2 are inner limiting sets so is K, their sum.

Suppose now that $\bar{K}$ contains a nondegenerate continuum. Then it contains an arc t. Since no subset of K is dense in itself, therefore there exists an interval t_1 of t which contains no point of K. The interval t_1 is a continuum of condensation of M and therefore, if H is the set of all emanation points of simple triods lying in M, t_1 is a subset of the closure of H (and indeed of the closure of $H \cdot t_1$). But this is contrary to Theorem 3.

[1] Cf. Besicovitch, A. S., *Fund. Math.*, **22**, 49–56 (1934).

[2] Throughout this paper space is supposed to be the plane.

[3] Whyburn, G. T., "Sets of Local Separating Points of a Continuum," *Bull. Amer. Math Soc.*, **39**, 97–100 (1933).

[4] Menger, Karl, "Grundzüge einer Theorie der Kurven", *Math. Annalen*, **95**, 272–306 (1925).

TORUS HOMOTOPY GROUPS*

BY RALPH H. FOX

DEPARTMENT OF MATHEMATICS, SYRACUSE UNIVERSITY

Communicated January 18, 1944

A promising line of attack on the unsolved problem of calculating the homotopy groups π_r, $r = 1, 2, \ldots$, of a topological space Y was initiated by J. H. C. Whitehead.[1] The results which he has obtained involve a certain "multiplication": for any pair of elements $\alpha \in \pi_m$, $\beta \in \pi_n$ the Whitehead product $\alpha \cdot \beta$ is defined and is an element of the group π_{m+n-1}. This multiplication can be compared with the group operation only when $m = n = 1$ and in this case $\alpha \cdot \beta = \alpha\beta\alpha^{-1}\beta^{-1}$.

In order to study this product I define here, for every dimension $r \geqslant 1$, a group τ_r which I call the r-dimensional torus homotopy group and which has the following properties:

1. Every homotopy group of dimension less than $r + 1$ can be mapped isomorphically into τ_r. (In fact most homotopy groups have many such isomorphisms.)

2. If $\gamma = \alpha \cdot \beta$, where $\alpha \in \pi_m$, $\beta \in \pi_n$, $\gamma \in \pi_{m+n-1}$ and $m + n - 1 < r + 1$, then isomorphisms $\pi_m \to \tau_r$, $\pi_n \to \tau_r$, $\pi_{m+n-1} \to \tau_r$ can be so chosen that $\alpha \to \bar{\alpha}$, $\beta \to \bar{\beta}$, $\gamma \to \bar{\gamma}$ and $\bar{\gamma} = \bar{\alpha}\bar{\beta}\bar{\alpha}^{-1}\bar{\beta}^{-1}$.

The groups $\tau_1, \tau_2, \tau_3, \ldots$ form a direct homomorphism system and the limit group τ has properties (1) and (2) with $r = \infty$. Thus all the homotopy groups and all the Whitehead products can be studied within one single group τ.

Let Y be a topological space and y_* one of its points. Denote by T_{r-1} the $(r - 1)$-dimensional torus whose coordinates are the $r - 1$ real numbers $x_1, x_2, \ldots, x_{r-1}$ (mod 1). The r-dimensional torus homotopy group $\tau_r(Y) = \tau_r(Y, y_*)$ is defined to be the fundamental group of the function space $Y^{T_{r-1}}$, using the mapping $y_* = y_*^{T_{r-1}}$ as base point. Thus an element of $\tau_r(Y)$ is represented by a continuous Y-valued function f of r real variables $x_0, x_1, \ldots, x_{r-1}$ (mod 1) which satisfies the condition

$$(*) \qquad f(x_0, x_1, \ldots, x_{r-1}) = y_* \text{ if } x_0 \equiv 0.$$

Denoting by $\mathfrak{L}_r$ the totality of such functions f, it is easy to see that two such functions f_0 and f_1 determine the same element of τ_r whenever there is a homotopy f_t, $0 \leq t \leq 1$, between them such that $f_t \in \mathfrak{L}_r$ for every t. The multiplication in τ_r is determined by the multiplication in $\mathfrak{L}_r$: $h = f \cdot g$ is defined by

$$h(x_0, x_1, \ldots, x_{r-1}) = \begin{cases} f(2x_0, x_1, \ldots, x_{r-1}) \text{ when } 0 \leq x_0 \leq 1/2, \\ g(2x_0 - 1, x_1, \ldots, x_{r-1}) \text{ when } 1/2 \leq x_0 \leq 1. \end{cases}$$

The discussion of the preceding paragraph parallels a discussion of the classical homotopy groups[2] which is presumedly known to the reader. It is sufficient here to recall that an element of $\pi_m(Y) = \pi_m(Y, y_*)$ is represented by a continuous Y-valued function f of m real variables $x_0, x_1, \ldots, x_{m-1}$ (mod 1) which satisfies the condition

(**) $\quad f(x_0, x_1, \ldots, x_{m-1}) = y_*$ if $x_i \equiv 0$ for some $i = 0, 1, \ldots, m-1$.

Let the totality of such functions f be denoted by $\mathfrak{F}_m$. If $m \leq r$ and A denotes a sequence $\{i_1, i_2, \ldots, i_{m-1}\}$, where $0 < i_1 < \ldots < i_{m-1} < r$, then, by assigning to every $f \in \mathfrak{F}_m$ that $g \in \mathfrak{L}_r$ which is defined by

$$g(x_0, x_1, \ldots, x_{r-1}) = f(x_0, x_{i_1}, x_{i_2}, \ldots, x_{i_{m-1}}),$$

a homomorphism $\omega_r^A\colon \alpha \rightarrow \alpha^A$ of $\pi_m(Y)$ into $\tau_r(Y)$ is set up. These homomorphisms can be shown to be isomorphisms, although the proof of this fact is by no means easy. There are exactly $\binom{r-1}{m-1}$ of these isomorphisms $\pi_m \rightarrow \tau_r$, and there is no reason for preferring any of them over the others. It is easily seen that τ_1 is identical with π_1, and that τ_2 is identical with Abe's[3] group κ_2.

A homomorphism Φ of τ_r onto τ_{r-1} is defined by assigning to every $f \in \mathfrak{L}_r$ that $g \in \mathfrak{L}_{r-1}$ which is defined by

$$g(x_0, x_1, \ldots, x_{r-2}) = f(x_0, x_1, \ldots, x_{r-2}, 0).$$

It can be shown that the subgroup τ_r' of τ_r which is generated by the subgroups $\omega_r^A(\pi_m)$, $2 \leq m \leq r'$, $A \subset \{1, 2, \ldots, r-1\}$, is the direct product $\mathrm{X}\omega_r^A(\pi_m)$ of these subgroups, that τ_r' is the nucleus of the homomorphism Φ and that τ_r is a split extension of the abelian subgroup τ_r' by τ_{r-1}.

In general the torus homotopy groups are non-abelian. In particular τ_r is non-abelian in the case that Y is the union of an m-sphere and an n-sphere $(m + n - 1 = r)$ which have exactly one point in common. Thus τ_r may be non-abelian for simply connected spaces if $r \geq 3$ and for spaces with abelian fundamental group if $r \geq 2$. This shows that not all of the "non-abelian character" of a space is expressed by its fundamental group. However if Y is a topological group its torus homotopy groups are all abelian.

The exact statement of how the Whitehead products are represented in τ_r as commutators is as follows: *If A has $m - 1$ indices and B has $n - 1$ indices, if $A \frown B$ is vacuous and if $\alpha \in \pi_m(Y)$ and $\beta \in \pi_n(Y)$ then $(\alpha \cdot \beta)^{A \smile B} = [\alpha^A, \beta^B]^{\epsilon\eta}$, where $[u, v]$ denotes the commutator $uvu^{-1}v^{-1}$; $\epsilon = (-1)^{n-1}$ and $\eta = (-1)^w$ where w is the number of instances of $i > j$ with $i \in A$ and $j \in B$.* It should be observed that it follows from previous discussion that $[\alpha^A, \beta^B] = 1$ if $A \frown B$ is not vacuous. By the theory of group-extensions

it can be shown that τ_r is determined by the groups $\pi_1, \pi_2, \ldots, \pi_r$ and the Whitehead products $\alpha \cdot \beta$, $\alpha \epsilon \pi_m$, $\beta \epsilon \pi_n$, $m + n - 1 < r + 1$.

By restating in terms of torus homotopy groups the combination of two of J. H. C. Whitehead's results[1] we find the following theorem: *Let K be a complex obtained from a complex K^* by removing the interior $\sigma - \dot{\sigma}$ of a principal n-dimensional simplex σ, where $n > 2$. The nucleus of the injection homomorphism $\tau_n(K) \to \tau_n(K^*)$ is precisely the invariant subgroup of $\tau_n(K)$ which is generated by the image of the injection homomorphism $\tau_n(\dot{\sigma}) \to \tau_n(K)$.* This reformulation seems to have more intuitive content and suggests more clearly than the original theorems an attack on the harder problem where $\dim \sigma > n$.

If Y_* is a subset of Y which contains y_* the relative torus homotopy group $\tau_r(Y \bmod Y_*, y_*)$, $r \geq 2$, may be defined. One starts with the collection $\mathfrak{L}_r(Y \bmod Y_*, y_*)$, of those continuous Y-valued functions f of the $r - 1$ real variables $x_0, x_1, \ldots, x_{r-2}$ (mod 1) and the real variable $0 \leq t \leq 1$ which satisfy the conditions

$$(***) \qquad \begin{cases} f(x_0, x_1, \ldots, x_{r-2}, t) \epsilon Y_* \text{ if } t = 0, \\ f(x_0, x_1, \ldots, x_{r-2}, t) = y_* \text{ if } t = 1 \text{ or if } x_0 \equiv 0. \end{cases}$$

From $\mathfrak{L}_r (Y \bmod Y_*, y_*)$ the group $\tau_r(Y \bmod Y_*, y_*)$ is obtained by the same procedure which led to the definition of $\tau_r(Y, y_*) = \tau_r(Y \bmod y_*, y_*)$. Just as in the absolute case there are isomorphisms ω_r^A: $\pi_m(Y \bmod Y_*) \to \tau_r(Y \bmod Y_*)$, where $A \subset \{1, 2, \ldots, r - 2\}$, and homomorphisms Φ: $\tau_r(Y \bmod Y_*) \to \tau_{r-1}(Y \bmod Y_*)$. The subgroup $\tau_r'(Y \bmod Y_*)$ of $\tau_r(Y \bmod Y_*)$ which is generated by the subgroups ω_r^A $(\pi_m(Y \bmod Y_*))$, $m \geq 3$, $A \subset \{1, 2, \ldots, r - 2\}$, is the direct product of these subgroups, τ' is the nucleus of Φ and τ is a split extension of τ_r' by τ_{r-1}

A homomorphism of $\tau_{r+1}(Y \bmod Y_*)$ into $\tau_r(Y_*)$ is defined by the transformation $f \to f \mid \{t = 0\}$; a homomorphism of $\tau_r(Y_*)$ into $\tau_r(Y)$ is defined by injection of Y_* into Y; a homomorphism of $\tau_{r+1}'(Y)$ into $\tau_{r+1}(Y \bmod Y_*)$ is defined by $f \to g$ where $g(x_0; x_1, \ldots, x_{r-1}; x_r) = f(x_0, x_A)$, $A \subset \{1, 2, \ldots, r\}$, $2 \leq m \leq r + 1$. In the resulting system of homomorphisms

$$\tau_{r+1}'(Y) \to \tau_{r+1}(Y \bmod Y_*) \to \tau_r(Y_*) \to \tau_r(Y)$$

it may be verified that the nucleus of any homomorphism is the image of the preceding homomorphism and that the nucleus of the first homomorphism is generated by the images of the injection homomorphisms $\pi_m(Y_*) \to \pi_m(Y)$, $2 \leq m \leq r + 1$. If Y is a fibre space over a space Z and Y_* is the fibre which contains y_* and whose image point is the base point z_* of Z then $\tau_r(Y \bmod Y_*, y_*) \approx \tau_r(Z, z_*)$ and the homomorphism system becomes

$$\tau_{r+1}'(Y) \to \tau_{r+1}(Z) \to \tau_r(Y_*) \to \tau_r(Y).$$

Any element of τ_r determines, together with the fundamental r-cycle of the antecedent space, a continuous r-dimensional cycle. In this fashion a homomorphism $\tau_r(Y)$ into the r-dimensional homology group $H_r(Y)$ of Y is defined. The image group is the group of spherical r-dimensional cycles, and the nucleus of this homomorphism obviously contains the commutator subgroup of τ_r. However, it is not true that the commutator subgroup is the nucleus. This is shown by the example of the 3-sphere $Y = S^3$ with $r = 4$. It is known that there is a non-trivial element α of $\pi_4(S^3)$, hence α^A is an element of the nucleus; on the other hand $\tau_4(S^3)$ is abelian because S^3 is a group manifold.

Elements of the nucleus which are not commutators may be constructed with the help of the homotopy groups of the rotation groups. If $\alpha \in \pi_n(Y)$ and ρ is an element of $\pi_m(R_{n-1})$, where R_{n-1} is the rotation group of the $(n-1)$-sphere then there is determined an element $\alpha^\rho \in \tau_{m+n}(Y)$. The element $(\alpha^\rho)^A(\alpha^{-1})^B$, where $A = \{1, 2, \ldots, m + n - 1\}$ and $B = \{1, 2, \ldots, n - 1\}$, belongs to the nucleus of the homomorphism $\tau_{m+n}(Y) \rightarrow H_{m+n}(Y)$. If α is the generator of $\pi_2(S^2)$ and ρ is the generator of $\pi_1(R_1)$ then $(\alpha^\rho)^A(\alpha^{-1})^B$ is the generator of $\pi_3(S^2)$. From the fact that the "Einhängung" of this element is non-trivial it can be shown that $(\alpha^\rho)^A(\alpha^{-1})^B$ is not a commutator.[4]

The torus homotopy groups have a rather obvious generalization which may turn out to be useful. Let μ be any mapping of T_{r-1} into Y and consider the fundamental group $\tau_r(Y, \mu)$ of the function space $Y^{T_{r-1}}$ using the mapping μ as base point. If $\mu = y_*$ the definition reduces to that of $\tau_r(Y, y_*)$. The group $\tau_r(Y, \mu)$ depends only on r, Y and the homotopy class of μ; particularly, to any element α of $\tau_r(Y, \mu)$ there is a group $\tau_{r+1}(Y, \alpha)$. The cycles which arise from the homomorphism of $\tau_r(Y, \mu)$ into $H_r(Y)$ need not be spherical cycles; for example, the fundamental cycle of the torus can occur.

A more detailed analysis of the torus homotopy groups will appear elsewhere.

* Presented to the American Mathematical Society under the title "The Complete Homotopy Group," abstracts 49-11-306(1943) and 51-1-49 (1945).

[1] *Ann. Math.*, **42**, 409–428 (1941) and *Proc. London Math. Soc.*, **45**, 243–327 (1939), §§10, 11.

[2] All homotopy groups are multiplicative in this paper.

[3] *Jap. Jour. Math.*, **16**, 169 (1940).

[4] In this and the succeeding paragraph there is a certain amount of overlapping with recent unpublished work by George W. Whitehead.

EXPERIMENTS ON SEXUAL ISOLATION IN DROSOPHILA. IV. MODIFICATION OF THE DEGREE OF ISOLATION BETWEEN DROSOPHILA PSEUDOOBSCURA AND DROSOPHILA PERSIMILIS AND OF SEXUAL PREFERENCES IN DROSOPHILA PROSALTANS

BY ERNST MAYR AND TH. DOBZHANSKY

THE AMERICAN MUSEUM NATURAL HISTORY AND COLUMBIA UNIVERSITY, NEW YORK

Communicated January 9, 1945

Sexually active animals find their potential mates and recognize them as belonging to the same species with the aid of stimuli that are but poorly known. In birds, it has been shown that species recognition may be either strictly innate or conditioned through experience (for a short summary of the literature see Cushing[1]). Conditioning seems to play a major rôle particularly in species with highly developed parental care. On the other hand, innate mechanisms control the recognition of potential mates of the same species in birds that are not raised by their own parents, such as cowbirds, parasitic cuckoos and megapodes. The same seems to be true for most of the lower vertebrates and invertebrates. It is, however, very little known to what extent the functioning of the innate patterns may be influenced by conditioning and by other extrinsic factors. The experiments described below were devised to explore the possibilities in this field.

Material and Methods.—For most of the experiments an orange-eyed mutant strain of *Drosophila pseudoobscura* Frolova descended from flies collected at Piñon Flats, San Jacinto Mountains, California, and a wild strain of *Drosophila persimilis* Dobzhansky and Epling from Stony Creek, north of the Sequoia National Park, California, were used. For some of the experiments strains of *Drosophila prosaltans* Duda from Chilpancingo and Zopilote Canyon, Mexico, and from Belem, Iporanga, and Bertioga, Brazil, were employed (concerning these strains see Dobzhansky and Streisinger[2]).

The two species, *D. pseudoobscura* and *D. persimilis*, are almost indistinguishable morphologically, although Reed, Williams and Chadwick[3] were able to discriminate the strains at their disposal with the aid of an ingeniously contrived ratio of thorax volume to the product of wing area times the cubed wing length. *D. persimilis* was formerly known as "*D. pseudoobscura* race B." The irrationality of this designation became progressively more clear with the accumulation of data showing that these species are distinct genetic systems; their separation is fully maintained in nature despite the broad overlapping of their distribution areas.[4] Too great an emphasis on morphological distinctions as species criteria leads to results that are plainly untenable. One would have to break living mankind into five species that are not at all isolated reproductively,[5] and yet consider as

single "species" some groups of clearly separate species of Drosophila.[6]

Except when stated otherwise, the experimental procedure was the same as described in the preceding parts of this series.[2,7] Batches of ten freshly hatched females of each of two species or strains to be tested were placed with ten males of one of these species in vials with food. After 4 to 7 days, depending on species and the particular experiment, the females were dissected and their sperm receptacles and spermathecae examined for sperm. The amounts of sperm present in an inseminated female vary greatly, particularly in heterogamic crosses: sometimes the ventral receptacle is tightly filled with sperm, sometimes loosely, and sometimes only a few moving spermatozoa are found. No attempt was made to record the various degrees of insemination; so long as any sperm was found, the fly was recorded as inseminated. In only a single *D. pseudoobscura* female, sperm was found in the spermatheca but none in the receptacle. The stock bottles were kept mostly at room temperature, and the experimental vials in an incubator at $24^1/_2$°C. Since the experiments lasted from May to November, "room temperature" varied considerably, and this may be a source of error in some of the experiments.

Sexual isolation between *D. pseudoobscura* and *D. persimilis* was first discovered by Lancefield[8] and subsequently studied by Boche,[9] but the data of the last-named investigator have never been published. We found the isolation to be much stronger when *D. pseudoobscura* than when *D. persimilis* males are used; it must, however, be noted that our experiments concern a single strain of each species, and that other strains may quite conceivably behave differently. Such differences between strains of *D. pseudoobscura* and *D. persimilis* with respect to sexual isolation from a third species, namely, *D. miranda*, are, indeed, known.[10]

Mixed Cultures.—Specific smells may be very important components of isolating mechanisms in animals with a highly developed olfactory sense. Experiments were, therefore, arranged to test whether or not there was a difference in the degree of sexual isolation between *D. pseudoobscura* and *D. persimilis* when these flies are raised in separate culture bottles or together in the same bottle. Fertilized females of the two species were placed together, but without males, in the same culture bottle, and transferred to fresh bottles at about 24-hour intervals. The larvae of both species grew up together in the same culture medium. The flies of the two species were separated after hatching before any copulation could occur. Sets of 10 females of both species were then confined with 10 males of one or the other species in vials with food, whereupon the females were dissected and their seminal receptacles were examined for sperm. As a control, similar tests were made using flies of the two species which developed in separate bottles.

It can be seen from table 1 that raising flies of the two species in the same culture medium does not lower the sexual isolation between them; curiously

enough, the results seem to indicate, if anything, the opposite. It probably occurs in nature not infrequently that larvae of the sympatric species *D. pseudoobscura* and *D. persimilis* grow up in the same food medium, and it is obviously of survival value to both species that the mixing of the smells of their larvae does not lead to lowering of isolation between the two species.

TABLE 1

NUMBER OF FEMALES DISSECTED (*n*) AND PER CENT CARRYING SPERM (%) IN CROSSES IN WHICH *D. pseudoobscura* AND *D. persimilis* FLIES WERE RAISED TOGETHER IN THE SAME BOTTLE OR IN DIFFERENT BOTTLES

MALE	RAISED	HOMOGAMIC *n*	HOMOGAMIC %	HETEROGAMIC *n*	HETEROGAMIC %	χ^2	ISOLATION INDEX
pseudoobscura	separately	82	87.8	83	9.6	100.7	0.80
pseudoobscura	together	145	89.6	140	0.7	226.6	0.98
persimilis	separately	127	63.8	119	22.7	42.5	0.48
persimilis	together (May–June)	41	73.2	41	2.4	43.6	0.94
persimilis	together (July)	82	76.5	89	32.9	31.2	0.40

Conditioning.—A set of *D. pseudoobscura* males was divided in two parts; some males were kept for 8–15 days in regular culture bottles with an excess of females of their own species ("pro-conditioned"), and others for the same length of time with females of *D. persimilis* ("counter-conditioned"). Similarly, some *D. persimilis* males were "pro-conditioned" and others were "counter-conditioned." Groups of 10 males were, then, confined with 10 freshly hatched females of each of the two species in vials with food; the females were dissected and examined for sperm. In "control" experiments freshly hatched males were confined with freshly hatched females. The results are summarized in table 2.

TABLE 2

NUMBER OF FEMALES DISSECTED (*n*) AND PER CENT CARRYING SPERM (%) IN CROSSES OF *D. pseudoobscura* AND *D. persimilis*

MALES	HOMOGAMIC *n*	HOMOGAMIC %	HETEROGAMIC *n*	HETEROGAMIC %	χ^2	ISOLATION INDEX
pseudoobscura control	82	87.8	83	9.6	100.7	0.80
pseudoobscura pro-conditioned	97	81.4	93	0.0	129.0	1.00
pseudoobscura counter-conditioned	115	88.7	122	1.6	182.7	0.96
persimilis control	127	63.8	119	22.7	42.5	0.48
persimilis pro-conditioned	32	56.3	37	13.5	13.9	0.69
persimilis counter-conditioned	47	87.2	52	38.5	25.0	0.39

The results are inconclusive as far as *D. pseudoobscura* males are concerned. There seems to be less isolation among the controls than among the counter-conditioned flies. However, the "control" experiment employed freshly hatched males while in the conditioning experiments males were 8–15 days old. Furthermore, most of the control experiments were made in May while the conditioning experiments were made in July. In

the case of *D. persimilis* males the conditioning appears to be effective. Males that had been conditioned with their own females show a higher isolation index (0.69) than the controls (0.48) or males conditioned with *D. pseudoobscura* females (0.39). The χ^2 of the difference between control and pro-conditioned flies is 2.08 (for two degrees of freedom $P > 0.20$); the χ^2 of the difference between control and counter-conditioned flies is 13.63 ($P < 0.01$). The effects of counter-conditioning are more significant than those of pro-conditioning.

Light.—Philip, Rendel, Spurway and Haldane[11] have stated that *D. subobscura*, a European relative of *D. pseudoobscura* and *D. persimilis*, does not mate in the absence of light, and that normal females of this species kick off mutant males with a yellow body color. Although morphological differences between *D. pseudoobscura* and *D. persimilis*, as well as those between strains of *D. prosaltans*, are very slight, the possibility that visual stimuli are involved in mate recognition is not excluded. To test this, vials were prepared containing two kinds of females and a single kind of males.

TABLE 3

MATE DISCRIMINATION IN THE LIGHT AND IN THE DARK

LIGHT OR DARK	FEMALES	MALES	HOMOGAMIC n	HOMOGAMIC %	HETEROGAMIC n	HETEROGAMIC %	ISOLATION INDEX
Light	pseudoobscura, persimilis	pseudoobscura	40	80.0	40	7.5	0.83
Dark	pseudoobscura, persimilis	pseudoobscura	60	80.0	69	2.9	0.93
Light	pseudoobscura, persimilis	persimilis	100	78.0	100	40.0	0.32
Dark	pseudoobscura, persimilis	persimilis	100	93.0	100	60.0	0.22
Light	prosaltans-A, prosaltans-D	prosaltans-A	70	82.9	65	3.1	0.93
Dark	prosaltans-A, prosaltans-D	prosaltans-A	69	46.4	68	1.5	0.94
Light	prosaltans-B, prosaltans-C	prosaltans-B	68	39.7	68	2.9	0.86
Dark	prosaltans-B, prosaltans-C	prosaltans-B	59	18.6	57	0.0	1.00

Some of the vials prepared on each of the days during which the experiments lasted were placed in an opaque box and the others on the top of the same box; the box was exposed to daylight but protected from direct sunlight. The temperature varied in the environment, but it was obviously very similar inside and outside the box. Females of the "dark series" were dissected soon after being removed from the box. The results are summarized in table 3; in this table the strains of *D. prosaltans* coming from Chilpancingo, Zopilote, Belem and Bertioga are denoted "prosaltans-A," B, C and D, respectively.

It is evident that in *D. pseudoobscura*, *D. persimilis* and *D. prosaltans* the mate discrimination is not greatly influenced by the presence or absence of light. In *D. prosaltans* the light has, however, an obvious influence on the total number of inseminations taking place within a given time interval; a significantly greater number of matings takes place in the vials exposed to light than in those kept in the dark. The data for *D. persimilis*,

if taken at their face value, would indicate an opposite effect of light, but the differences observed are in need of confirmation.

The Role of the Wings.—When a courting male of *D. pseudoobscura* or *D. persimilis* pursues a female he spreads and vibrates his wings. The pitch of vibration may be correlated with the wing surface[9] which is larger in the latter than in the former species. If females exercise the choice, it is possible that they are helped in recognition of the males by the pitch of the wing vibration. If so, the females should have difficulties in recognizing wingless males, and the isolation index should drop. Actually, the opposite happened when wingless *D. persimilis* males were confined with winged *D. persimilis* and *D. pseudoobscura* females—the isolation index became higher (table 4). However, the point when around 50 per cent of the females were inseminated was not reached after 4–5 days' exposure, as with normal males, but only after 8 days. Females seem to recognize the species of wingless males as readily as of normal ones, but either succeed better in avoiding insemination by not conspecific wingless males or are less easily excited into a receptive state.

TABLE 4

INSEMINATION RECORDS OF *D. persimilis* AND *D. pseudoobscura* FEMALES BY WINGLESS *D. persimilis* MALES

HOMOGAMIC n	HOMOGAMIC %	HETEROGAMIC n	HETEROGAMIC %	χ^2	ISOLATION INDEX
78	65.4	70	12.8	42.8	0.67

Experiments with wingless females and normal males resulted in isolation indices which are practically identical with the control experiments. This is important if considered in conjunction with the observation that non-receptive females often flick off with their wings males which attempt to mount them. Non-receptive wingless females seem to be equally capable of avoiding males.

Sexual excitement.—Dobzhansky and Koller,[10] working with *D. pseudoobscura* and *D. miranda*, obtained an indication that males aged in the absence of females are less efficient in discriminating between their own and foreign females than males pro-conditioned with their own females. If significant, this result may be due either to sexual excitement of the males aged without females or to the pro-conditioning of the other group of males. We have kept males and females of *D. pseudoobscura*, *D. persimilis*, and of four strains of *D. prosaltans* in isolation from individuals of the opposite sex but with abundant food for approximately seven days, whereupon these fully mature flies were placed together in the same vial, always avoiding etherization of the flies. As in the earlier experiments, one kind of males and two kinds of females were placed in each vial. The difference between this technique and that of Dobzhansky and Koller[10] lies in that in the experi-

ments under consideration both males and females were aged in isolation, while in the latter experiments only the males were so aged. Courting and copulating pairs can be seen in the vials within a few minutes after the flies are placed together, and in from one to four hours approximately half of all females are found to be inseminated. If freshly hatched flies are used, it takes from four to five days for half of the females to become inseminated, and relatively few copulating pairs are seen in the vials at any one time. The results of the experiments are summarized in table 5. The

TABLE 5

MATE DISCRIMINATION IN INDIVIDUALS AGED IN ISOLATION FROM THE OPPOSITE SEX

FEMALES	MALES	HOMOGAMIC n	HOMOGAMIC %	HETEROGAMIC n	HETEROGAMIC %	χ^2	ISOLATION INDEX
pseudoobscura, persimilis	pseudoobscura	40	70.0	37	5.4	20.53	0.85
pseudoobscura, persimilis	persimilis	43	67.4	56	7.1	26.65	0.81
prosaltans-A, prosaltans-D	prosaltans-A	63	93.6	74	23.0	30.50	0.61
prosaltans-B, prosaltans-E	prosaltans-E	38	2.6	43	67.4	22.63	−0.92

Chilpancingo, Zopilote, Bertioga and Iporanga strains of *D. prosaltans* are referred to in this table as "prosaltans-A," B, D and E, respectively. Frequencies of the homogamic and heterogamic fertilizations in mixtures of *D. pseudoobscura* and *D. persimilis* shown in table 2 may be taken as control values for comparison with the data in table 5, although these experiments have not been performed simultaneously. For insemination records in *D. prosaltans* see Dobzhansky and Streisinger.[2]

Examination of table 5 shows that aging in the absence of individuals of the opposite sex fails to change the degree of sexual isolation when *D. pseudoobscura* males are used; with *D. persimilis* males, such aging leads even to a strengthening of the isolation, although more data are needed to establish this point. Aged prosaltans-A (Chilpancingo) males gave a somewhat lower isolation index than was obtained with males placed together with their prospective mates shortly after their hatching from the pupae; males of prosaltans-E (Iporanga strain) prefer B (Zopilote) females to their own, and this preference seems to be enhanced by aging in the absence of mates.

Temperature.—Flies raised at room temperature were placed in vials soon after their hatching from pupae, and the vials with the flies were kept at $24^1/_2$°, 21°, 18° and $16^1/_2$°C. for as long as necessary to obtain insemination of about half of the females. This takes 4–5 days at the higher and 7–9 days at the lower temperatures. The results are summarized in table 6. "Prosaltans-A," B, C and D are, in this table, the Chilpancingo, Zopilote, Belem and Bertioga strains of *D. prosaltans*, respectively.

The behavior of *D. pseudoobscura* and *D. prosaltans* flies is about the same at all the temperatures tried. *D. persimilis* shows clear sexual isola-

tion from *D. pseudoobscura* at the higher temperatures, but at 18° and 16½° *D. persimilis* males seem to discriminate against females of their own species in favor of those of *D. pseudoobscura*. This is particularly astonishing because *D. persimilis* is, on the whole, confined in nature to cooler habitats than *D. pseudoobscura*. It may be that females of *D. persimilis* become sexually receptive only very slowly at temperatures of 18°C. and lower, so that most *D. persimilis* females in the low temperature experiments were simply unavailable for insemination. To test this possibility, *D. persimilis* females and males, and *D. pseudoobscura* females, were aged for 10 days at 16½°C., and placed together, at the same temperature, for about 24

TABLE 6

MATE DISCRIMINATION AT DIFFERENT TEMPERATURES

$t°$	FEMALES	MALES	HOMOGAMIC n	HOMOGAMIC %	HETEROGAMIC n	HETEROGAMIC %	χ^2	ISOLATION INDEX
24½°	pseudoobscura, persimilis	pseudoobscura	30	83.3	28	3.6	20.4	0.92
18°	pseudoobscura, persimilis	pseudoobscura	21	85.7	18	0.0	15.4	1.00
16½°	pseudoobscura, persimilis	pseudoobscura	42	92.9	40	12.5	24.2	0.76
24½°	pseudoobscura, persimilis	persimilis	65	93.8	64	39.1	14.6	0.41
21°	pseudoobscura, persimilis	persimilis	56	53.6	63	12.7	15.4	0.62
18°	pseudoobscura, persimilis	persimilis	21	4.8	20	55.0	8.7	−0.84
16½°	pseudoobscura, persimilis	persimilis	86	32.6	90	52.2	4.0	−0.23
24½°	prosaltans-A, prosaltans-C	prosaltans-A	59	86.4	58	8.6	37.8	0.82
24½°	prosaltans-B, prosaltans-D	prosaltans-B	58	74.1	58	8.6	30.1	0.79
16½°	prosaltans-A, prosaltans-C	prosaltans-A	77	90.9	75	2.3	62.3	0.94
16½°	prosaltans-B, prosaltans-C	prosaltans-B	84	44.0	85	4.7	14.6	0.81

hours. Dissection of the females showed that 59.4% of the 106 *D. persimilis*, and 44.6% of the 92 *pseudoobscura* females were inseminated. This gives a non-significant positive isolation index of 0.14. Sexual isolation between *D. pseudoobscura* and *D. persimilis* appears, then, to be weaker at lower than at higher temperatures if males of *D. persimilis* are used.

Acknowledgments.—We are indebted to Dr. M. Demerec, the Director of the Biological Laboratory of the Long Island Biological Association at Cold Spring Harbor, for his hospitality and his interest in our work. The assistance of Miss Irene Markreich and of Mr. George Streisinger in conducting the experiments is gratefully acknowledged.

[1] Cushing, John E., Jr., *The Condor*, **43**, 233–236 (1941).

[2] Dobzhansky, Th., and Streisinger, G., these PROCEEDINGS, **30**, 340–345 (1944).

[3] Reed, S. C., Williams, C. M., and Chadwick, L. E., *Genetics*, **27**, 349–361 (1942).

[4] Dobzhansky, Th., and Epling, C., *Carnegie Inst. Washington, Publ.* **554**, 1–46 (1944).

[5] Gates, R. R., *Amer. Jour. Phys. Anthrop.*, 2 (n.s.), 279–292 (1944).

[6] Sturtevant, A. H., *Univ. Texas Publ.*, **4213**, 5–51 (1942).

[7] Dobzhansky, Th., and Mayr, E., these PROCEEDINGS, **30**, 238–244 (1944).

[8] Lancefield, D. E., *Zeits. ind. Abst. Vererbungsl.*, **52**, 287–317 (1929).

[9] Boche, R. D., in Dobzhansky's *Genetics and the Origin of Species*, New York, 1937.

[10] Dobzhansky, Th., and Koller, P.Ch., *Biol. Zentralb.*, **58**, 589–607 (1938).

[11] Philip, U., Rendel, J. M., Spurway, H., and Haldane, J. B. S., *Nature*, **154**, 260–262 (1944).

DOMINANCE MODIFICATION AND PHYSIOLOGICAL EFFECTS OF GENES*

BY L. C. DUNN AND S. GLUECKSOHN-SCHOENHEIMER

DEPARTMENT OF ZOÖLOGY, COLUMBIA UNIVERSITY

Communicated January 15, 1945

Several years ago[1] there was found in the house mouse a mutation with several striking effects including absence or shortening of the tail, absence or abnormality of one or both kidneys, absence of external anus and genital aperture and abnormalities of other parts of the urogenital system. In the stock in which it occurred this whole syndrome of effects behaved as a unit and showed simple segregation from normal. The mutation acted as a lethal, all homozygotes *Sd Sd* being tailless and dying shortly after birth, all showing imperforate anus and absence of both kidneys. Heterozygotes showed a lesser expression of these defects, having short tails, and less severe urogenital malformations. The mutation in the original stock therefore acted as a dominant in respect to its effect on the tail, as recessive or nearly so in its lethal effect, and as incompletely dominant in its effect on urogenital development.

When the *Sd* mutation was removed from the stock in which it originally occurred and was transferred by a series of successive backcrosses to another inbred normal stock, the tail length of heterozygotes progressively decreased until after five backcross generations nearly all were tailless, while the viability of the heterozygotes decreased, due to the greater effect of *Sd* on the urogenital system.[1] About 90 per cent of all *Sd*+ animals at birth had abnormal kidneys.[2] The dominance of *Sd* on tail length appeared to have been increased by the genetic constitution of the new stock while the lethal effect appeared also to have become partially dominant. There was no evidence of necessary connection between the effect upon tail length and upon viability.

Since the above observations were published we have transferred the *Sd* mutation to two other normal-tailed inbred stocks by repeated backcrossing. In one of these stocks (identified as *m*) the tail length of heterozygotes increased, and the proportion of tailless animals among the heterozygotes decreased. In the F_1, BC_1 and BC_2 generations the cross of *Sd*+ by normal *m* produced 142 normal, 90 short-tailed and 25 tailless; while in BC_3 and BC_4 the comparable figures are 39 normal, 23 short and no tailless. The

totals, 181 normal and 138 Sd+ $X^{2-5.26}_{p=0.2}$, indicate a *lowered* viability of *Sd*+ associated with *increasing* tail length. *This was probably due to increased* severity of the urogenital lesions, since of four heterozygotes dissected at birth, three had urogenital abnormalities.

When backcrossed to the other inbred stock, a normal albino stock known as CF (Carworth Farms), changes in the opposite direction occurred in the *Sd* heterozygotes; that is, the tails became shorter, the proportion of tailless heterozygotes *increased*, and the relative viability of the heterozygotes *increased*. The figures are F_1–BC_2: 59 normal, 14 short, 53 tailless; BC_3–BC_6: 26 normal, 7 short, 18 tailless. The totals of 85 ++ 92 *Sd*+ indicate no excess mortality of short and tailless *Sd*+. Forty-three of these heterozygotes were dissected and in all cases the urogenital system was normal. Thus the CF genetic constitution, which *increased* the severity of the tail defect, eliminated the deleterious effect of *Sd*+ on viability and on the urogenital defects chiefly responsible for the viability effects of *Sd*. The conclusion is obvious that the several effects of *Sd* upon the heterozygotes are modified by *different* genetic factors.

Since this is so, the conclusions of Fisher and Holt[4] concerning dominance modification of *Sd* will have to be examined critically. Fisher and Holt similarly outcrossed *Sd*+ animals obtained from this laboratory to other stocks in the Cambridge University laboratory, and set up selection lines in one of which the tail length of heterozygotes was increased, while in the other, tail length remained short. The viability of heterozygotes in the longer-tailed line improved. Fisher and Holt supposed that the chief factor in viability was tail length, so that natural selection for viability aided the selection for longer tails, while in the negative line it was "always acting in opposition, in that mice most defective in the development of the caudal vertebrae will also, on the whole, be most defective in other respects." This assumption, as our observations show, is entirely unwarranted. They assume also that as the tail length of the heterozygotes increases "so one might expect the chance of homozygotes becoming less abnormal, with a consequent lengthening of life, to increase." This is based on the same reasoning as that above and is supported only by the occurrence in the line selected for tail length, of two animals diagnosed as homozygotes (*Sd Sd*), one of which lived for 22 days and the other for 72 hours. Since tailless heterozygotes with internal lesions as severe as those of the longer lived assumed homozygote have been described by Gluecksohn-Schoenheimer,[2] the diagnosis of homozygosity rests on the occurrence of imperforate anus and cloaca. These defects have been found recently in rare cases among heterozygotes also, so it is possible that Fisher and Holt were dealing with an extreme heterozygote.

There is no doubt that the tail length of heterozygous *Sd*+ mice is readily modified by other genetic factors, which thus may be said to modify the

dominance of *Sd* in its effect upon the tail. But these other factors, as our results show, may change the dominance of *Sd* in its other effects in ways opposite to that in which the tail expression is affected.

In order to derive some meaning from such observations for Fisher's general theory of the origin of dominance in evolution, it would be necessary to specify whether modifiers of *Sd* would be selected because of their effect on tail length or because of their effect on urogenital development. In view of the evident relation between the urogenital expression and viability there can be little doubt that modifiers which tend to make the urogenital effect of *Sd* recessive would be selected. But these same genetic constitutions may, as we have seen, increase the dominance of *Sd* in respect to tail length. It can be concluded that the evolutionary significance of factors affecting dominance can be properly assessed only when their physiological effects are known.

* The observations reported were made with the aid of a grant from the Josiah Macy, Jr., Foundation.

[1] Dunn, L. C., Gluecksohn-Schoenheimer, S., and Bryson, V., "A New Mutation in the Mouse Affecting Spinal Column and Urogenital System," *Jour. Heredity*, **31**, 343–348 (1940).

[2] Gluecksohn-Schoenheimer, S., "The Morphological Manifestations of a Dominant Mutation in Mice Affecting Tail and Urogenital System," *Genetics*, **28**, 341–348 (1943).

[3] Dunn, L. C., "Changes in the Degree of Dominance of Factors Affecting Tail-Length in the House Mouse," *The American Naturalist*, **76**, 552–569 (1942).

[4] Fisher, R. A., and Holt, S. B., "The Experimental Modification of Dominance in Danforth's Short-Tailed Mutant Mice," *Ann. Eugenics*, **12**, 102–120 (1944).

RELATIVE SENSITIVITY OF CHROMOSOMES TO NEUTRONS AND X-RAYS. III. COMPARISON OF CARCINOMA AND LYMPHOSARCOMA IN THE RAT

By A. Marshak and Muriel Bradley*

Radiation Laboratory, University of California, and Department of Radiology, University of California Medical School

Communicated January 15, 1945

In previous studies it was shown that the relative response of chromosomes of the "resting" nucleus to neutrons and x-rays (n/x) may be used as means for identifying physiological stages of the nucleus which cannot otherwise be recognized.[1, 2] By applying such an analysis to different tumors in the same species it is possible to determine whether the nuclei and chromosomes of the cell types so analyzed are in fact the same, as might be expected from existing theories in genetics and embryology, or

whether they differ. We wish to report here the results of an analysis of the Walker 256 carcinoma and a lymphosarcoma of the rat.

TABLE 1

Walker Carcinoma 256

TIME IN HOURS	X-RAYS DOSE IN r UNITS	TOTAL ANA-PHASES	% NORMAL	COR-RECTED % NORMAL	NEUTRONS DOSE IN n UNITS	TOTAL ANA-PHASES	% NORMAL	COR-RECTED % NORMAL
3	47	569	61.5	66.5	8	708	62.7	67.7
	99	526	37.1	42.1	8.9	711	60.2	65.2
	143	568	30.1	35.1	12.5	584	48.9	53.9
	180	533	13.7	18.7	17.8	552	30.4	35.4
	240	539	7.9	12.9	27	613	17.3	22.3
	286	537	2.4	7.4	38	563	7.7	12.7
	...	...	...	...	41	740	7.4	12.4
8	50	606	69.1	74.1	9	518	61.7	66.7
	107	532	48.9	53.9	18	552	40.2	45.2
	164	681	36.3	41.3	24	616	30.2	35.2
	228	734	26.3	31.3	30	579	24.2	29.2
	277	692	17.1	22.1	44	580	15.7	20.7
	344	647	8.5	13.5	56	834	9.2	14.2
12	48	714	66.1	71.1	7	715	58.8	63.8
	98	608	47.6	52.6	21	545	31.6	36.6
	149	637	38.5	43.5	22	645	27.0	32.0
	192	846	28.9	33.9	37	547	15.2	20.2
	300	859	15.9	20.9	40	733	9.4	14.4
	337	797	11.4	16.4	54	516	4.8	9.8
	354	666	7.4	12.4	...	...	...	...
	450	655	4.6	9.6	...	...	...	...
18	120	671	63.6	68.6	6	565	75.9	80.9
	254	928	34.0	39.0	20	664	46.8	51.8
	392	536	24.8	29.8	36	575	26.1	31.1
	462	557	20.5	25.5	48	686	13.1	18.1
	588	527	10.1	15.1	49	597	11.9	16.9
	684	623	4.9	9.9	62.9	722	7.9	12.9
	...	...	...	...	63.7	572	6.6	11.6
	...	...	...	...	74.4	633	4.6	9.6
	...	...	...	...	81.7	675	1.7	6.7
24	95	583	69.6	74.6	18	627	55.8	60.8
	120	606	77.2	82.2	24	648	42.3	47.3
	239	627	44.6	49.6	43	584	26.9	31.9
	392	863	30.0	35.0	55	607	18.8	23.8
	462	577	27.9	32.9	92	658	5.9	10.9
	588	505	20.5	25.5	104	744	3.1	8.1
	784	637	10.8	15.8	...	...	...	...

The animals carrying the tumors were treated with x-rays or with neutrons by the same methods and apparatus previously described.[2] The methods for fixing the tissue, preparing the smears and making the counts have also been presented.

The data obtained from counts of normal and abnormal anaphases are presented in tables 1 and 2. For plotting the curves (Figs. 1, 2, 3 and 4) each per cent normal anaphases was corrected by adding to it the average per cent abnormal anaphases of all controls for that tumor. The correction for the carcinoma was 5.0%, the average from 12 control animals; for the lymphosarcoma it was 12.1%, the average from 7 controls. The per cent

TABLE 2

LYMPHOSARCOMA

	X-RAYS				NEUTRONS			
TIME IN HOURS	DOSE IN r UNITS	TOTAL ANA-PHASES	% NORMAL	COR-RECTED % NORMAL	DOSE IN n UNITS	TOTAL ANA-PHASES	% NORMAL	COR-RECTED % NORMAL
3	51	543	63.9	76.0	8	519	62.6	74.7
	97	538	43.1	55.2	16	518	47.8	59.9
	148	578	31.3	43.4	22	506	32.4	44.5
	201	513	17.6	29.7	34	504	18.6	30.7
	272	529	6.4	18.5	51	503	8.9	21.0
	341	543	2.8	14.9	58	521	2.6	14.7
8	96	545	56.9	69.0	10	601	68.5	80.6
	166	544	45.4	57.5	20.5	520	50.9	63.0
	226	526	28.5	40.6	31	540	37.2	49.3
	301	525	21.5	33.6	41	511	27.8	39.9
	365	524	15.1	27.2	51	516	18.6	30.7
	424	516	9.5	21.6	63	521	10.1	22.2
12	55	553	66.7	78.8	8	559	73.9	86.0
	117	555	52.0	64.1	21	517	45.6	57.7
	197	520	33.8	45.9	39	518	28.2	40.3
	274	524	25.2	37.3	45	524	23.6	35.7
	361	536	17.9	30.0	57	539	14.1	26.2
	417	543	12.3	24.4	67	522	12.4	24.5
18	95	556	55.8	67.9	...	...	...	...
	189	540	44.8	56.9	...	...	...	...
	295	634	24.9	37.0	...	...	...	...
	384	553	14.3	26.4	...	...	...	...
	484	539	8.7	20.8	...	...	...	...
	566	524	5.5	17.6	...	...	...	...
24	94	572	64.4	76.5	10	629	72.9	85.0
	206	596	44.9	57.0	22.5	550	58.2	70.3
	307	531	34.8	46.9	42	530	44.5	56.6
	414	526	27.3	39.4	51	546	37.0	49.1
	539	590	15.4	27.5	84	516	18.0	30.1
	637	545	12.1	24.2	89	540	13.5	25.6

abnormal anaphases in unirradiated animals is considerably greater in the lymphosarcoma than in any other type of tumor studied so far. When plotted semi-logarithmically the per cent normal anaphases for each time interval following irradiation fall on a straight line, i.e., the data fit an exponential curve. The points were all fitted graphically. The mean deviation of the points so fitted varied from 0.7 to 3.7%, with only

one curve giving a mean deviation greater than 3%, so the fit may be considered good. Table 3 gives the slopes of the curves, the mean deviation of the points about each curve, and the ratio (n/x) of the slopes obtained with neutrons to those obtained with x-rays.

The slopes of the curves for the lymphosarcoma are consistently smaller than the corresponding slopes for the carcinoma for both x-rays and neu-

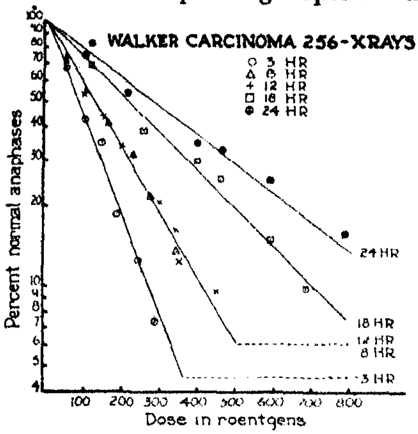


FIGURE 1

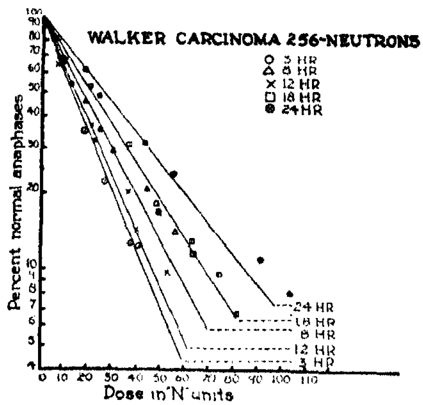


FIGURE 2

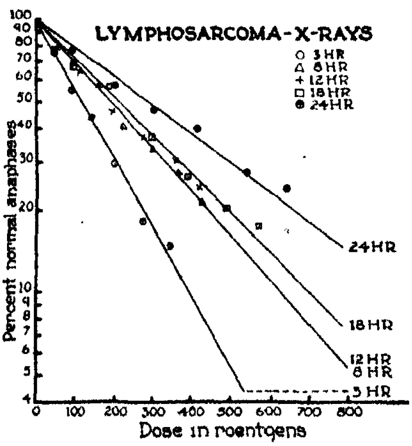


FIGURE 3

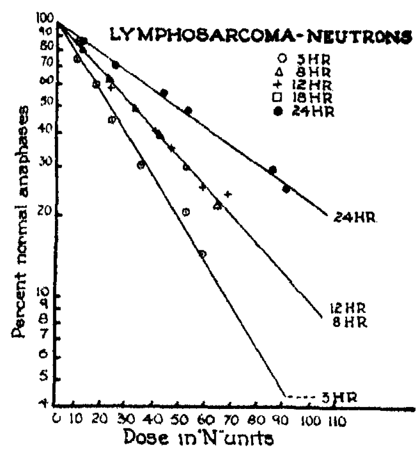


FIGURE 4

trons at all time intervals except at 24 hours. Such a result might be expected if the chromosomes of the lymphosarcoma were smaller than those of the carcinoma. However, the chromosomes of the cells studied showed no very obvious differences in size, although it was found impossible to make any accurate measurements. We cannot, therefore, offer any explanation for the observed differences in slopes.

Although there is a progressive decrease in the slopes observed following x-rays or neutrons in both tumors at 8, 18 and 24 hours after irradiation, the slopes at 12 hours are anomalous in this respect. In the lymphosarcoma there is no significant difference between the slopes at 8 and 12 hours, which is also true for the carcinoma treated with x-rays. The latter tumor when treated with neutrons, however, gives a steeper slope at 12 than at 8 or 18 hours. It is interesting to note that analysis of chromosome response in the plant *Vicia faba* shows no difference in x-ray slopes at 12 and 18 hours and no significant difference in 8- and 12-hour neutron slopes. In the mouse lymphoma there is a marked rise in the slope for both x-rays and neutrons at 12 hours. If slopes of survival curves for either x-rays or neutrons are plotted as a function of time after irradiation, in all the tissues thus analyzed the slope for the 12-hour interval appears either as part of a plateau or as a peak in the curve. Obviously the 12-hour stage is a critical one in all the chromosomes analyzed so far.

TABLE 3

	TIME IN HOURS	$k_x \times 10^{-3}$	d_x	$k_n \times 10^{-3}$	d_n	n/x
Walker 256	3	8.45	1.8	50.81	2.0	6.01
carcinoma	8	5.52	1.5	40.76	2.7	7.28
	12	5.58	2.3	49.27	2.9	8.83
	18	3.25	2.0	33.62	1.1	10.34
	24	2.49	3.7	26.93	2.0	10.81
Lympho-	3	5.86	1.5	33.80	1.8	5.77
sarcoma	8	3.65	1.5	22.73	0.7	6.23
	12	3.66	2.2	22.91	1.9	6.26
	18	3.32	1.3	...	...	...
	24	2.44	2.2	14.62	1.6	5.99

k_x = slopes following treatment with x-rays, k_n = slopes following treatment with neutrons, d_x = deviation of points from x-ray curves, d_n = deviation of points from neutron curves, n/x = ratio of the slopes obtained with neutrons to those obtained with x-rays.

The ratio of neutron to x-ray efficiency (n/x) is approximately 6 at 3 hours for both tumors, as it has been for chromosomes of all species of plants and animals studied so far.[3] Beyond this point all similarity between the n/x ratios for the two tumors ceases. While there is a progressive increase in n/x to 10.8 at 24 hours in the carcinoma, in the lymphosarcoma it remains at approximately 6 for all the periods studied.† However, the ratios for the lymphosarcoma are strikingly similar to those observed in a mouse lymphoma, for which the ratios were 5.8, 5.9, 8.8, 5.9, and 5.9 for the intervals 3, 8, 12, 18 and 24 hours.[2] In other words, the chromosomes of the rat lymphosarcoma in their reaction to x-rays and neutrons resemble those of the mouse lymphoma much more than they do chromosomes of another tumor of the rat.‡ For reasons previously given,[1, 2] the differences in re-

sponse of the chromosomes cannot be ascribed to differences in the extra-chromosomal (cytoplasmic) conditions of the histologically different types of cells involved. Likewise the theory advanced by other investigators[4, 5] which postulates chromosome "breakage" followed by "healing" of the broken ends, cannot account for results such as these.[1] We must conclude, therefore, that there are physiological differences in the chromosomes of two different tissues (lymphoid and epithelial) of the same species.

Theories concerning cell differentiation during the early stages of ontogeny and of "de-differentiation" in carcinogenesis commonly imply that the genic constitution of the cells remains constant. By implication it is also inferred that the chromosomes also remain unaltered. The experiments reported here indicate that such assumptions are entirely unwarranted. Possible changes in the chromosomes during embryogeny or carcinogenesis may be detected by the methods described here.

Summary.—1. Chromosomes of the Walker 256 rat carcinoma are more sensitive to x-rays and neutrons than those of a rat lymphosarcoma in all parts of the resting stage studied except at 24 hours prior to anaphase.

2. In both these tumors, as well as in a previously studied mouse lymphoma and the plant *Vicia faba*, the mitotic phase 12 hours before anaphase appears to be a critical one.

3. The ratio of neutron to x-ray efficiency, n/x, which was previously shown to be an index of physiological activity of the chromosomes, is approximately 6 at all time intervals studied in the lymphosarcoma, whereas in the carcinoma n/x increases progressively from 6.0 at 3 hours to 10.8 at 24 hours. From this it is concluded that there are physiological differences in the chromosomes of two different tissues in the same species.

4. Chromosome response of the rat lymphosarcoma to x-rays and neutrons is more like that of the mouse lymphoma than that of the rat carcinoma.

5. Results obtained here indicate that assumptions that the chromosomes remain unaltered during ontogeny of normal cells and during de-differentiation in carcinogenesis are unwarranted.

We wish to express our gratitude to Dr. Robert S. Stone and Dr. Earl R. Miller of the Department of Radiology for the interest they have shown in this research and for facilitating arrangements for its execution. We take this opportunity to thank the Department of Botany of the University of California for the use of equipment and facilities essential to this work.

* Fellow of the Finney-Howell Research Foundation.

Part of this research was supported by grants from the Christine Breon and Frank and Blanche Wolf Funds administered by the Medical Research Committee of the University of California Medical School.

† A breakdown in the cyclotron made it impossible to obtain the data needed for the 18-hour neutron curve. However, the argument presented here will not be altered by whatever value of n/x may be obtained for this interval.

‡ Although the mouse lymphoma and rat lymphosarcoma do show similarity in response as determined by the ratio n/x, there is a difference between the two in another respect. When abnormalities are plotted as a function of time after irradiation, the mouse lymphoma chromosomes show a peak in sensitivity at 12 hours as well as at 3 hours, which does not appear in the lymphosarcoma.

[1] Marshak, A., *Proc. Nat. Acad. Sci.*, **28**, 29–35 (1942).

[2] Marshak, A., *Radiology*, **39**, 621–626 (1942).

[3] Marshak, A., *Proc. Soc. Exp. Biol. Med.*, **41**, 176–180 (1939).

[4] Gustafsson, A., *Hereditas*, **23**, 281–335 (1937).

[5] Sax, K., *Genetics*, **26**, 418–425 (1941).

PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 31 March 15, 1945 Number 3

ON THE GENETIC CONTROL OF MUTABILITY IN MAIZE

By M. M. Rhoades

Department of Botany, Columbia University

Communicated February 3, 1945

The *Dt* gene in maize is of more than usual interest because of its striking ability to increase the mutability of the recessive *a* allele which is normally highly stable.[1, 2] The locus of *a* is in chromosome 3; data were published in 1941 proving that *Dt* lies in chromosome 9, but its precise location in the linkage map was uncertain. The present paper is chiefly concerned with (1) the accurate location of *Dt*, (2) the regional distribution of crossing-over in the short arm of chromosome 9 and the location of the known mutant genes in this chromosome, and (3) a consideration of a paper by Goldschmidt[3] in which is offered an alternative explanation to the results interpreted by the author as the greatly increased mutability of the recessive *a* allele.

(1) According to the summary in Emerson, Beadle and Fraser,[4] the linear order and intervening map distances in chromosome 9 are as follows: yg_2 19 *C* 3 *sh* 15 *bp* 15 *wx* 12 v_1. The yg_2 locus is close to the end of the short arm with the other loci occupying more proximal positions. Four point back-cross data involving the yellow green-2 (yg_2), shrunken endosperm (*sh*), waxy endosperm (*wx*) and the *Dt* loci are presented in table 1. The *Dt* allele has no known effect other than to raise the mutation rate of the *a* allele; the classification for *Dt* and *dt* therefore rests upon the presence or absence of mutant areas showing the *A* phenotype. The data in table 1 place *Dt* seven cross-over units to the left of yg_2. Creighton and McClintock[5] and McClintock[6] have shown that yg_2 is very close to the terminal knob on the end of the short arm. Creighton found only 1.5 per cent crossing-over between yg_2 and the knob. The location of *Dt* seven units to the left of yg_2, which is only 1.5 units from the knob terminating the end of the chromosome, would seem an inconsistency. However, the cross-over value found by Creighton for the knob-yg_2 region is undoubtedly too low, since she worked with chromosomes 9, one of which possessed a large terminal knob while the other had a minute knob. It has been observed

that asynapsis frequently occurs with terminal heteromorphic knobs; this asynapsis should result in lower cross-over values in distal regions. Further, her low percentage of recombination was based on a relatively small population of 261 individuals, which would permit a large sampling error.

(2) The location of *Dt* at the extreme left end of the short arm of chromosome 9 is of interest in connection with McClintock's[6] recent investigations on mutations and deficiencies. The short arm of this chromosome at pachytene has about 20 chromomeres. Those chromomeres adjacent to the centromere are more deeply staining than the distal ones. In an elegant series of experiments she localized the yg_2 locus in a distal portion of the first chromomere. Joining this ultimate chromomere to the terminal knob is a thin, lightly staining chromatin thread to which she was able to assign the locus of a gene affecting chlorophyll development. Disregarding the heterochromatic terminal knob, the only part of chromosome 9 to the left of the yg_2 locus is the threadlike strand connecting the ulti-

TABLE 1

FOUR POINT BACK-CROSS LINKAGE DATA INVOLVING THE *Dt* GENE

$$\frac{Dt \overset{(1)}{\ } Yg \overset{(2)}{\ } Sh \overset{(3)}{\ } Wx}{dt \quad yg \quad sh \quad wx} \times dt\ yg\ sh\ wx$$

(0)	(0)	(1)	(1)	(2)	(2)	(3)	(3)	(1-2)	(1-2)	(2-3)	(2-3)	(1-3)
Dt	*dt*	*Dt*	*dt*	*Dt*	*dt*	*Dt*	*dt*	*Dt*	*dt*	*Dt*	*dt*	*dt*
Yg	*yg*	*yg*	*Yg*	*Yg*	*yg*	*Yg*	*yg*	*yg*	*Yg*	*Yg*	*yg*	*Yg*
Sh	*sh*	*sh*	*Sh*	*sh*	*Sh*	*Sh*	*sh*	*Sh*	*sh*	*sh*	*Sh*	*Sh*
Wx	*wx*	*wx*	*Wx*	*wx*	*Wx*	*wx*	*Wx*	*Wx*	*wx*	*Wx*	*wx*	*wx*
880	672	88	99	225	309	251	196	3	4	12	13	3

Recombination percentages: *Dt*-*Yg* 7.2%
Yg-*Sh* 20.6%
Sh-*Wx* 17.2%

mate chromomere and the knob. *Dt* is seven cross-over units to the left of yg_2; unless it is situated within the terminal knob (a possibility which cannot be wholly dismissed even though the knobs are believed to be composed of genetically inert heterochromatin), it must lie somewhere on the threadlike strand. The total length of the genetic map of the short arm of chromosome 9 is approximately 60 units. Cytologically this short arm has approximately 20 chromomeres. Seven of the 60 cross-over units lie to the left of the ultimate chromomere. This is a much greater proportion than should occur if the amount of crossing-over per unit of length was of the same order along the short arm and suggests that a disproportionately high amount occurs in the distal portion of the short arm of chromosome 9.

It is perhaps significant that the *Dt* gene, which exerts such a profound influence on the mutability of another gene, is located near to or in the heterochromatic knob terminating the end of the short arm of chromo-

some 9. The mechanisms through which the *Dt* gene alters the mutability of *a* are unknown, but on certain hypotheses the effect of *Dt* would be modified if it were interstitially rather than terminally situated.

The location of *Dt* in the short arm of chromosome 9 is in no way surprising, but it might be pertinent to mention the distribution of mutant loci in this chromosome. The long arm of 9 is approximately twice the length of the short arm, yet of the thirty-odd mutant genes belonging to this chromosome not one has been located definitely in the long arm, although virescent-1 may fall here.[7] The failure to find a proportionate share of mutant loci in the long arm may reasonably be attributed to the existence of duplications.[8] These duplications may exist within the long arm of 9 itself or may be found in certain of the other chromosomes. A similar condition possibly exists for chromosome 5. The two arms are approximately equal in length, yet the great majority of the mutant genes have been found to lie in the slightly longer arm. With the exception of chromosome 8, which is as yet sparsely populated with mutant genes, such a marked localization of genes to restricted portions of the chromosome as occurs in chromosomes 9 and 5 does not appear to hold for the remaining chromosomes, although more information on this point is needed. That duplications do exist in maize is attested by the occurrence of occasional bivalents at M I in haploid plants. The existence of duplications is of interest in connection with the problem of the origin of maize. It has been held that maize is an ancient amphidiploid resulting from the cross of two five-chromosome species. The notion that five is the basic number rather than ten has been entertained ever since two close relatives, Coix and Sorghum, were found to have species with a haploid set of five chromosomes.

(3) In a recent paper, Goldschmidt[8] described a situation in *Drosophila melanogaster* in which factor interaction gives a variety of phenotypes resembling in some degree at least the array of types found by Demerec[9] for the unstable miniature alleles of *Drosophila virilis*. The $bran^{dp}$ allele of the arc locus in chromosome 2 has no phenotypic effect unless the recessive allele, svr^{poi}, at the silver locus in the X chromosome is present. Flies lacking the $bran^{dp}$ allele but homozygous for svr^{poi} have pointed wings, while flies with $bran^{dp}$ and svr^{poi} exhibit a considerable diversity of phenotypes. The majority have pointed wings like those of svr^{poi}, the remainder have one wing pointed and one truncated or else have "transitions from a pointed to a truncated wing in all conditions of asymmetry down to a symmetrical, truncated (dumpy-like) wing." Goldschmidt believes the $bran^{dp}$ allele to be one which acts near the threshold between the pointed wing and the dumpy wing phenotypes and holds that the range of phenotypes may be satisfactorily accounted for by epistasis near the threshold value and by consideration of the develop-

mental physiology of the wing. Goldschmidt states that the phenotypes found by Demerec for the unstable *mt-c* allele resemble those found in $bran^{dp}$ svr^{poi} flies and suggests that the unstable gene hypothesis be abandoned. He argues that the *mt-c* allele is not unstable but is influenced by another gene, similar to $bran^{dp}$, which has no effect alone but which interacts with *mt* in the direction of normal wing development. He reports another arc allele, $bran^{r}$, which gives an array of phenotypes resembling those produced by the unstable *mt-a* allele in *virilis*.

In the belief that he has satisfactorily accounted for the variety of phenotypes found for the unstable miniature alleles in *virilis* by a hypothesis which does not require the high mutation rate postulated by Demerec, he next considers variegation in plants. Reviewing the *a-Dt* situation in maize reported by the author, he suggests that the variegation found here can also be accounted for by his hypothesis of factor interaction, epistasis and threshold values. Specifically he cites the *a* allele as comparable to svr^{poi}, while *Dt* would parallel $bran^{dp}$. To account for mutations of *a* to *A* in sporogenous tissue, giving rise in many cases to anthers with half of the pollen carrying the *a* allele and the other half possessing the dominant *A* allele (derived by mutation), he assumes that the mutations of *a* to *A*, yielding self-colored seeds and plants, are not changes of *a* to *A*, but are due to the introduction through outcrossing of a new allele of *Dt* equal to *A* in its ability to produce anthocyanin in the aleurone, plant and pericarp tissues. This new allele of *Dt* is therefore comparable to the $bran^{r}$ allele. Goldschmidt states that no published results disprove his hypothesis. In this he errs, because several decisive observations which negate his interpretation have been published. First, the location of the *a* allele is in chromosome 3 while *Dt* is in chromosome 9. The newly arisen mutations of *a* to *A* show the same linkage relations with genes in chromosome 3 as do standard *A* alleles. On Goldschmidt's hypothesis the new color-producing genes should exhibit linkage with loci in chromosome 9 since *Dt* lies in that chromosome. Second, when a mutation of *a* to *A* occurs in a cell of *a a Dt Dt* constitution, the genotype of that cell and its descendents is *A a Dt Dt*. On his hypothesis the constitution should be *a a Dt* Dt^{A} (where Dt^{A} represents the new *Dt* allele equal to *A* in its ability to form anthocyanins). Third, Goldschmidt postulates the existence of a new allele of *Dt* which produces the same phenotypic effect as does *A*. At least five different alleles at the *a* locus have been obtained by mutation of *a* in the presence of *Dt*. On his hypothesis there must be five different *Dt* alleles, yet the *Dt* gene comes from a single source. They could arise only by mutation, so his hypothesis would merely shift the mutability from the *a* allele to the *Dt* allele, and he would still have to contend with the high mutability of the *Dt* allele.

Interesting though Goldschmidt's studies are on factor interaction in

Drosophila melanogaster, and however closely their phenotypic expression may simulate those produced by unstable genes, they can hardly be considered as negating the validity of the mutable gene hypothesis.

Summary.—Data are presented showing that the *Dt* allele, which profoundly affects the mutability of the recessive *a* allele, is located at the end of the short arm of chromosome 9. The regional distribution of crossing-over in the short arm of chromosome 9 and the location of the known mutant genes in chromosome 9 are considered. An examination of the published data on the genetic control of the mutation rate of the recessive *a* allele reveals that it cannot be accounted for in terms of factor interaction, epistasis and threshold values as Goldschmidt suggests.

[1] Rhoades, M. M., *Genetics*, **23**, 377–397 (1938).

[2] Rhoades, M. M., *Cold Spring Harbor Symposia on Quantitative Biology*, **9**, 138–144 (1941).

[3] Goldschmidt, Richard, *Proc. Nat. Acad. Sci.*, **29**, 203–206 (1943).

[4] Emerson, R. A., Beadle, G. W., and Fraser, A. C., *Cornell Univ. Agric. Expt. Station Memoir*, **180** (1935).

[5] Creighton, H. B., and McClintock, B., *Proc. Nat. Acad. Sci.*, **21**, 148–150 (1935).

[6] McClintock, B., *Genetics*, **29**, 478–502 (1944).

[7] Burnham, C. R., *Ibid.*, **19**, 430–447 (1934).

[8] Anderson, E. G., *Ibid.*, **23**, 307–313 (1938).

[9] Demerec, M., *Cold Spring Harbor Symposia on Quantitative Biology*, **9**, 145–150 (1941).

*MAINTENANCE AND INCREASE OF A GENETIC CHARACTER BY A SUBSTRATE-CYTOPLASMIC INTERACTION IN THE ABSENCE OF THE SPECIFIC GENE**

BY S. SPIEGELMAN, CARL C. LINDEGREN AND GERTRUDE LINDEGREN

DEPARTMENT OF BACTERIOLOGY AND IMMUNOLOGY, WASHINGTON UNIVERSITY SCHOOL OF MEDICINE AND THE HENRY SHAW SCHOOL OF BOTANY, WASHINGTON UNIVERSITY

Communicated January 29, 1945

Introduction.—Our previous studies[1,2] on the production of galactozymase by *Saccharomyces cerevisiae* revealed that in certain diploid strains its appearance in the cell was effected by a direct interaction between the cytoplasm and galactose. In certain haploid strains, mutation was apparently required before adaptation to galactose fermentation occurred. Study[3] of other yeast types supported the view that mutation in the haplophase was one of the mechanisms of adaptation. Thus, it was possible to adapt the normally diploid *Schizosaccharomyces pombe* (which had previously been reported[4] as unadaptable) to galactose fermentation by inoculating a galactose medium with a heavy suspension of haploid spores.

Information on the relation of specific genes to adaptation was pro-

vided by an analysis of the genetics of adaptation to melibiose fermentation.[6] In this investigation progenies of a hybrid between *Saccharomyces cerevisiae* and *Saccharomyces carlsbergensis* were used. *S. carlsbergensis* and all its haploid segregants can adapt to melibiose fermentation; *S. cerevisiae* and its haploid progeny cannot. Data obtained from 175 progenies of the interspecific and of related hybrids were consistent with the view that *S. carlsbergensis* is homozygous for two pairs of dominant genes, any one of which is capable of producing the adaptive enzyme.

The experiments with adaptation to both galactose and melibiose led to the conclusion that the gene initiates the synthesis of the adaptive enzyme, providing the specific substrate is present. The data showed that, within wide limits, the amount of adaptive enzyme in the cytoplasm depends on an interaction between the cytoplasm and the substrate. These results raised the following questions:

1. If synthesis has been initiated, can the substrate-cytoplasmic interaction maintain the enzyme indefinitely in the cytoplasm in the absence of the specific gene?

2. If some enzyme is present, can synthesis of additional enzyme occur in the absence of the specific gene necessary to initiate its synthesis?

The opportunity for answering these questions was provided by analysis of certain progenies of the *S. cerevisiae* by *S. carlsbergensis* pedigree.[5] Both questions were answered in the affirmative. Not only can the cells increase their ability to ferment melibiose in the absence of the gene but this character has been maintained in these cells for over 1000 generations.

Methods.—A. Matings and Isolation of Haploid Segregants: The general methods for inducing ascospore formation, their dissection and subsequent treatment for the production of hybrids and isolation of haploid strains have been described elsewhere.[6, 7, 8, 9]. For purpose of later discussion, the steps involved in the formation of a hybrid and the analysis of its haploid segregants are listed:

1. Mating by mixture of two haploid strains of proper mating types.
2. Induction of sporulation.
3. Dissection of 4-spored asci.
4. Planting of the four spores separately on agar.
5. Testing of the resulting clones for the character.

B. Test for Ability to Ferment Melibiose: All clones were characterized according to their ability to ferment melibiose by the standard inverted tube technique. These tests were run in duplicate and held for a period of more than four weeks before being discarded as negative. All these tests were checked by Warburg manometric methods, in which the rate of CO_2 evolution in nitrogen by a suspension of the cells in $M/15$ KH_2PO_4 was measured after the melibiose had been added from a sidearm. The nitrogen

used to displace the air in measurements on anaerobic CO_2 production was passed over hot copper to remove any traces of oxygen. All measurements were taken at 30.2°C. and the vessels were shaken at a rate of 100 oscillations per minute over a 7-cm. arc. The basic medium and carbohydrate, as well as their treatment and preparation, were the same as those used previously.[5]

Experimental.—In the previous[5] experiments on adaptability to melibiose fermentation, the cells came into contact with melibiose for the first time in the test for adaptability, and the first four steps (see *Methods*) were performed without the addition of melibiose. In the present experiments the effect of performing all five steps in the presence of melibiose (3%) was tested. The reason for this procedure lies in the fact that a cellular adaptive enzyme is stabilized by keeping the cells in contact with the specific substrate.

TABLE 1

EFFECT OF MELIBIOSE ON PHENOTYPIC CHARACTERS OF SEGREGANTS FROM DIPLOIDS FORMED BY MATINGS IN ITS PRESENCE AND ABSENCE. + INDICATES ABILITY TO FERMENT MELIBIOSE, − INABILITY. ALL SPORES COME FROM A (+ X −) CROSS. (SEE TEXT FOR FURTHER DETAILS)

MATING, SPORULATION AND PLANTING IN PRESENCE OF MELIBIOSE					MATING, SPORULATION AND PLANTING IN ABSENCE OF MELIBIOSE				
ASCUS NO.	SPORES A	B	C	D	ASCUS NO.	SPORES A	B	C	D
1	+	+	+	+	8	+	+	−	−
2	+	+	+	+	9	−	+	−	+
3	+	+	+	+	10	+	+	−	−
4	+	+	+	+	11	+	+	−	−
5	+	+	+	+	12	−	+	−	+
6	+	+	+	+	13	−	−	+	+
7	+	+	−	−	14	+	−	−	+
					15	+	+	−	−
					16	+	−	+	−
					17	−	−	+	+

A hybrid was obtained by mating an adaptable haplophase clone carrying a single *S. carlsbergensis* gene controlling adaptation, to a haplophase clone of *S. cerevisiae* which carried the allele controlling non-adaptability. These heterozygous diploid hybrids are all adaptable, for the mel + gene is dominant. Each 4-spored ascus from a heterozygous diploid cell yields two adaptable and two unadaptable haplophase cultures.

Table 1 summarizes the results of these experiments. Asci 1–15 originated from mating a pair of mel+/mel− haploids while 16 and 17 originated from mating a different pair of mel+/mel− haploids. Melibiose was present in the substrate in all five steps in the formation and dissection of asci 1–7, inclusive. Asci 10–17, inclusive, were formed in the usual way, without melibiose. In handling asci 8 and 9 the agar in which the spores were planted contained melibiose (step 4, *Methods*) although melibiose was absent in the first 3 steps.

It is evident from table 1 that all asci formed in the complete absence of melibiose give the typical 1 : 1 ratio characteristic of a heterozygous hybrid segregating a single pair of genes. These results agree with those reported previously[5] on equivalent crosses. On the other hand, with the exception of ascus No. 7, identical heterozygotes treated with melibiose yielded four adaptable spores from each ascus.

The results obtained without melibiose prove that only two spores of each tetrad in asci 1–6, inclusive, contain the specific gene responsible for adaptation to melibiose fermentation. Despite this, all four spores from these tetrads produced haplophase cultures which fermented melibiose.

Since all steps were carried out in the presence of melibiose, selection of adaptable mutants from haploids originally unable to ferment melibiose might have occurred. Step 4 particularly is open to such criticism. However, several specific facts rule out this possibility: (1) During the testing of many segregants from *S. cerevisiae*, all of which are negative, no mutation to an adaptable type has been observed whether melibiose was present or not. (2) The same is true of negative haploids from heterozygous hybrids. No mutations to the adaptable type have been seen among these no matter how often they have been transferred through melibiose media. (3) Asci 8 and 9 whose segregants were planted on melibiose, yielded the standard 1 : 1 ratio.

More conclusive evidence on this point was obtained by the following experiments. Presumably, the cultures from two spores of each tetrad (from the first six asci) were able to ferment melibiose only due to the presence of the adaptive enzyme in the cytoplasm. Consequently, it is to be expected that removal of the melibiose would lead not only to the disappearance of fermentability in all cases, but to an eventual loss of readaptability in two of every four cultures arising from each of the first six asci. To exclude the complication of mutations away from adaptability[3] only non-dividing cultures suspended in $M/15$ KH_2PO_4 were used. The 24 adapted haplophase cultures originating from the first six asci were all grown in the basic culture medium containing melibiose as the sole carbohydrate source. Using sterile precautions the cells were washed free of the medium with $M/15$ KH_2PO_4 and resuspended in sterile $M/15$ KH_2PO_4 to make suspensions containing approximately 8 mg. dry weight of yeast per cc. Each suspension was divided into two equal parts, and sufficient melibiose was added to one portion to make a 4% solution. This was used as a control to test the stability of the enzyme in the presence of the substrate under the experimental conditions. The other portion of each suspension did not receive any melibiose. The flasks were then shaken continuously at 28°C. and cells were removed at intervals to test them for the ability to ferment melibiose.

None of the control suspensions containing melibiose showed loss of the

ability to ferment this sugar within the experimental period. On the other hand, all of the suspensions without substrate showed decreases within 24 hrs. in their rates of anaerobic CO_2 evolution when melibiose was added. In varying periods of time, ranging from 7–20 days, all of these suspensions lost the ability to evolve significant amounts of CO_2 anaerobically on immediate contact with melibiose. Twenty-four hours after a suspension showed insignificant rates of CO_2 evolution in the presence of melibiose (i.e., $Q^N_{CO_2}$ values of less than unity) a sample was removed and incubated with melibiose aerobically at 30.2°C. to test for readaptability. At the same time its ability to ferment glucose was also examined. This was done to avoid testing cells whose physiological condition was seriously impaired by the long vigorous shaking in the phosphate buffer without substrate. With cells unable to ferment glucose the inability to readapt to melibiose would be difficult to interpret. Three suspensions of the original 24 were eliminated on this basis. The data on the asci producing four testable segregants are given in table 2. The removal of melibiose and its stabiliz-

TABLE 2

READAPTABILITY OF SPORES OBTAINED BY MATINGS IN PRESENCE OF MELIBIOSE AFTER HAVING LOST ALL ADAPTIVE ENZYMES. + INDICATES READAPTABILITY, – INABILITY

ASCUS NO.	SPORES A	B	C	D
1	+	–	+	–
2	–	–	+	+
4	+	+	–	–
6	+	+	–	–

ing influence leads to the reappearance of the expected Mendelian ratios. These results give further support to the view that only two spores in each of these tetrads carried the mel+ gene.

In addition, data collected at the same time show that synthesis of additional enzyme can occur in the absence of the specific gene necessary to initiate the synthesis. After allowing all suspensions to fall to low $Q^N_{CO_2}$ values (between 1.8 and 10.1), portions were removed and incubated with melibiose and regeneration of activity followed at intervals by measuring the rate of anaerobic CO_2 evolution. The results on those haploid segregants which subsequently lost the ability to adapt are recorded in table 3. It is seen that in all cases, marked increases in activity were obtained. It may be noted here without going into detail that no significant difference in the rate of increase in enzyme activity could be established between the cells carrying the non-fermenting allele (mel–) and those which possessed the mel+ gene. As far as can be determined it appears that the rate of enzyme regeneration is critically determined by the enzyme and substrate content of the cytoplasm at the outset of the incubation period.

All the strains listed in table 3 were carried in standard media with melibiose and were tested at weekly intervals. At the end of three months they could all ferment melibiose at a rate equal to or greater than the original rate. This period is equivalent to over 1000 cell generations. It is thus evident that melibiose can maintain the enzyme in the cytoplasm of these cells while active division is going on for long periods of time.

Discussion.—These results appear to indicate that melibiose can obscure the nuclear hereditary mechanism because it is the specific substrate for a cytoplasmic enzyme whose activity level depends critically on the amount of substrate available. The fact that adaptability was lost on the removal of substrate proves that a directed mutation under the influence of melibiose is not involved.

An explanation of the results obtained, in terms of enzymes and their specific substrates, may be briefly summarized as follows: By performing the mating in the presence of melibiose, the cytoplasm of the haploid gamete carrying the mel+ gene is packed with the melibiose fermenting enzyme. Since both copulating haploids contribute cytoplasm equally to the zygote, it contains the enzyme. Because sporulation occurs in the presence of melibiose, the enzyme molecules are stabilized and possibly increased in amount, since the pre-sporulation period is characterized by growth and considerable storage. Each of the four haploid segregants derives its cytoplasm from the diploid hybrid and it follows that each will have enzyme molecules in its cytoplasm no matter what its genetic constitution. Finally, the enzyme molecules are stabilized even in the descendents of the spores which do not have the mel+ gene by keeping the cells in contact with melibiose. From this point of view, both maintenance and increase of the enzyme activity can be effected by a direct cytoplasmic interaction with the substrate in the absence of the specific gene. Therefore, as far as melibiozymase is concerned the function of the mel+ gene would be limited to the initiation of its synthesis.

TABLE 3

$Q^{N}_{CO_2}$ VALUES AFTER AEROBIC INCUBATION WITH MELIBIOSE OF STRAINS WHICH EVENTUALLY LOST THEIR ABILITY TO ADAPT

STRAIN	HOURS IN CONTACT WITH MELIBIOSE 0	12	24	48
1B	5.1	40	96	123
1D	2.4	26	109	114
2A	10.1	39	86	136
2B	6.3	46	73	101
4C	5.0	69	160	170
4D	4.2	29	91	134
6C	1.8	34	84	141
6D	4.8	42	121	130

The experiments reported here do not rule out the possibility of a more direct genic involvement than is implied in the explanation proposed above. Thus, e.g., the data could be satisfied by assuming the existence of a third gene in the carlsbergensis genome in addition to the two mel+ genes already described. This third gene, however, would have to possess the following properties or their equivalents:

(a) It can be stimulated to functional activity in the presence of either one of the mel+ genes and the substrate, but not by melibiose alone.

(b) Once activated it can remain functional in the absence of both mel+ genes, providing melibiose is present.

(c) It becomes irreversibly deactivated if melibiose is removed and all the melibiozymase exhausted from the cytoplasm.

Various modifications of this type of genic theory may be easily constructed. However, on the basis of the available evidence, the most plausible explanation of the data which can be advanced at present is in terms of substrate stabilized enzymes and the effect of their specific substrates on their maintenance.

The fact that enzymes in the cytoplasm can be influenced by their specific substrates provides a mechanism by which some of the phenomena of cytoplasmic inheritance can be explained. Tronsitory cytoplasmic effects could result from residual enzyme molecules which are not stabilized by the proper substrate. Lindegren[10] explains the degeneration of hybrids showing heterosis on this basis. The importance of cytoplasmic units in inheritance has been emphasized by the work of Winge and Laustsen,[11] Sonneborn[12] and Rhoades[13] among others and the speculations of Wright[14] and Darlington.[15] The adaptive enzyme, melibiozymase is a self-perpetuating cytoplasmic entity but differs from Darlington's "plasmagene" in being gene initiated and substrate stabilized.

Since substrate can evoke enzymatic activity previously non-detectable, it is conceivable that this phenomenon may also play a role in cellular differentiation.

Summary.—The effect of melibiose on the inheritance of the ability to ferment this sugar has been studied. In the absence of melibiose, a 1:1 ratio of fermentors to non-fermentors is exhibited by the four haploid segregants from an ascus. This is typical of a heterozygous diploid and agrees with the known genetic background of the hybrid employed. When the segregation occurs in the presence of melibiose *all four* segregants can ferment the sugar. All four haploids and the clones derived from them can maintain this ability indefinitely if kept in contact with the substrate. When, however, the melibiose is removed, only two out of the four can re-adapt to its fermentation. These data are explained in terms of adaptive enzymes and their stabilization by their specific substrates in the absence

of the gene; transfer of the character from one cell generation to the next is apparently effected by the enzyme molecules contained in the cytoplasm.

* We would like to express our appreciation for a grant from Anheuser-Busch, Inc., St. Louis, Mo., and to the Department of Zoölogy of Washington University for the laboratory facilities placed at the dispòsal of one of us (S. S.).

[1] Spiegelman, S., Lindegren, C. C., and Hedgecock, L., *Proc. Nat. Acad. Sci.*, **30**, 13–25 (1944).

[2] Spiegelman, S., and Lindegren, C. C., *Ann. Mo. Botan. Garden*, **31**, 219–233 (1944).

[3] Spiegelman, S., and Lindegren, C. C., *Jour. Bact.* (in press) (1944).

[4] Armstrong, E. F., *Proc. Roy. Soc. London*, **76B**, 600–605 (1905).

[5] Lindegren, C. C., Spiegelman, S., and Lindegren, G., *Proc. Nat. Acad. Sci.*, **30**, 346–352 (1944).

[6] Lindegren, C. C., and Lindegren, G., *Ann. Mo. Botan. Garden*, **30**, 453–469 (1943).

[7] Lindegren, C. C., and Lindegren, G., *Proc. Nat. Acad. Sci.*, **29**, 306–308 (1943).

[8] Lindegren, C. C., and Lindegren, G., *Bot. Gaz.*, **105**, 304–316 (1944).

[9] Lindegren, C. C., and Lindegren, G., *Ann. Mo. Botan. Garden*, **31**, 203–216 (1944).

[10] Lindegren, C. C. (in MS.).

[11] Winge, Ö., and Laustsen, Ö., *Compt. Rend. Trav. Lab. Carlsb. Ser. Physiol.*, **23**, 17–38 (1940).

[12] Sonneborn, T. M., *Proc. Nat. Acad. Sci.*, **29**, 338–343 (1943).

[13] Rhoades, M. M., *Ibid.*, **29**, 327–329 (1944).

[14] Wright, S., *Physiol. Rev.*, **21**, 487–527 (1941).

[15] Darlington, C. O., *Nature*, **154**, 164–169 (1944).

ON SURFACE AREA

BY TIBOR RADO

INSTITUTE FOR ADVANCED STUDY AND THE OHIO STATE UNIVERSITY

Communicated February 8, 1945

1. We shall be concerned in this note with Fréchet surfaces of the type of the 2-cell.[7,8] Such a surface S admits of a representation of the form $S: x = x(u, v)$, $y = y(u, v)$, $z = z(u, v)$, $(u, v) \in Q$, where Q is the unit square $0 \leqq u \leqq 1$, $0 \leqq v \leqq 1$, and $x(u, v)$, $y(u, v)$, $z(u, v)$ are continuous in Q. The Lebesgue area of S will be denoted[8] by $A(S)$. We introduce a quantity $a(S)$, to be termed the lower area of S, using conceptions in the theory of continuous transformations in the plane.[3,6] These conceptions will be briefly reviewed presently.

2. Let D be a bounded domain (connected open set) in the uv plane. A bounded continuous transformation T, from D into a $\xi\eta$ plane, is given in the form $T: \xi = \xi(u, v)$, $\eta = \eta(u, v)$, $(u, v) \in D$, where $\xi(u, v)$, $\eta(u, v)$ are bounded continuous functions in D. It is not assumed that T is bi-unique. With each point (ξ, η) there is associated[8] its essential multiplicity $\kappa(\xi, \eta,$

T, D) with respect to T and D. For each point (ξ, η) we consider the set $T^{-1}(\xi, \eta)$ in D. A maximal model continuum for (ξ, η) under T in D is a compact component of $T^{-1}(\xi, \eta)$. Such a continuum may or may not be essential.[3] We define a subset E^* of D as follows: E^* is the sum of all the essential maximal model continua, under T in D, corresponding to all the points (ξ, η) of the $\xi\eta$ plane. The set E^* depends upon both T and D. Let $\bar{D}$ be a subdomain of D. We denote by $\varphi(\bar{D})$ the measure of the set of those points (ξ, η) where $\kappa(\xi, \eta, T, \bar{D}) \pm 0$. In particular, $\varphi(\bar{D})$ is thus defined for rectangles in D. The derivative[3] of this rectangle function, if it exists at a point (u, v), will be denoted by $D_e(u, v)$. Here and in the sequel, a subscript e refers to the fact that the quantity involved is defined in terms of the essential multiplicity κ. We define further a subset N of D as follows. A point $(u_0, v_0) \epsilon D$ belongs to N if and only if the following conditions hold: (i) $(u_0, v_0) \epsilon E^*$; (ii) there exists an open set G such that $(u_0, v_0) \epsilon GD$, and the set $G - (u_0, v_0)$ contains no point of any essential maximal model continuum of the point $T(u_0, v_0)$. It follows that there exist simple closed curves C of arbitrarily small diameter, such that (u_0, v_0) is interior to C and the topological index of the point $T(u_0, v_0)$ with respect to the image of C under T is different from zero.[3] It follows further that this index is independent of C if the diameter of C is sufficiently small. We denote this index by $i_e(u_0, v_0)$. For points (u, v) not in N, we put $i_e(u, v) = 0$. If $D_e(u, v)$ exists at a point (u, v), then the quantity $J_e(u, v) = i_e(u, v)D_e(u, v)$ will be termed the essential generalized Jacobian.[3] By an oriented rectangle r we mean a rectangle with sides parallel to the u and v axes, respectively. r° denotes the interior of r. If the rectangle function $|T(r^\circ E^*)|$ is absolutely continuous[3] in D, then the transformation T will be said to be eAC in D (essentially absolutely continuous in D). If the essential multiplicity function $\kappa(\xi, \eta, T, D)$ is summable, then T will be said to be eBV in D (of essential bounded variation in D).[3,6]

3. Given a surface S as in 1, we introduce the three transformations (projections upon the coordinate planes):

$$\begin{aligned} T_x&: y = y(u, v),\ z = z(u, v),\ (u, v) \epsilon Q^0, \\ T_y&: z = z(u, v),\ x = x(u, v),\ (u, v) \epsilon Q^0, \\ T_z&: x = x(u, v),\ y = y(u, v),\ (u, v) \epsilon Q^0. \end{aligned}$$

The given representation of S will be termed[6] eBV if T_x, T_y, T_z are eBV in Q^0, and eAC if T_x, T_y, T_z are eAC in Q^0. The essential generalized Jacobians of T_x, T_y, T_z will be denoted by X_e, Y_e, Z_e, respectively. We shall denote the arithmetic square root of $X_e^2 + Y_e^2 + Z_e^2$ by W_e. If W_e exists almost everywhere in D and is summable in Q^0, then its integral over Q^0 will be denoted by I_e. If the first partial derivatives of $x(u, v)$, $y(u, v)$, $z(u, v)$ happen to exist at a point (u, v), then X, Y, Z will denote the Jacobians of T_x, T_y, T_z in the usual sense. W will then denote the arithmetic square

root of $X^2 + Y^2 + Z^2$. If W exists almost everywhere in Q^0 and is summable in Q^0, then its integral over Q^0 will be denoted by I. If $\bar{D}$ is a domain in Q^0, then we shall denote by $g_x(\bar{D})$ the integral of $\kappa(y, z, t_x, \bar{D})$ if this integral exists in the Lebesgue sense; otherwise we put $g_x(\bar{D}) = +\infty$. The quantities $g_y(\bar{D})$, $g_z(\bar{D})$ are defined in a similar manner. Finally, we denote by $g(\bar{D})$ the arithmetic square root of $g_x(\bar{D})^2 + g_y(\bar{D})^2 + g_z(\bar{D})^2$. Let now $\bar{D}_1, \bar{D}_2, \ldots$ be any finite or infinite system of disjoint domains in D. The least upper bound of the summation $g(\bar{D}_1) + g(\bar{D}_2) + \ldots$, for all possible systems $\bar{D}_1, \bar{D}_2, \ldots$, is defined as the lower area $a(S)$ of S.

4. If we restrict, in the definition of $a(S)$, the domains $\bar{D}_1, \bar{D}_2, \ldots$ by permitting only simply connected domains, then we obtain a quantity that has been studied by Reichelderfer,[6] and has been called by him the essential area of S. Clearly $a(S)$ is an upward revision of the essential area, which itself is an upward revision of other and analogously defined lower areas. The lower area $a(S)$, defined above, is always less than or equal to the Lebesgue area $A(S)$. Generally speaking, most substantial results concerning the Lebesgue area were achieved in cases where it could be shown that some kind of a lower area agreed, under the special circumstances involved, with $A(S)$. The lower area $a(S)$, introduced above, is the largest lower area studied so far, and thus it may be expected that it will agree with $A(S)$ under more general conditions. On the other hand, it is conceivable that in the process of upward revision we lose desirable properties as compared with previously used smaller lower areas. The outcome, as it appears at this time, is summarized in the following statements.

5. THEOREM. *Given S as in 1, suppose that $A(S) < +\infty$. Then the given representation of S is eBV in Q^0, the essential generalized Jacobians X_e, Y_e, Z_e exist almost everywhere in Q^0, and $I_e \leq A(S)$ (cf. 3). The sign of equality holds if and only if the given representation is eAC in Q^0.*

6. THEOREM. *Given S as in 1, suppose that $A(S) < +\infty$, and suppose also that the first partial derivatives of $x(u, v)$, $y(u, v)$, $z(u, v)$ exist almost everywhere in Q^0. Then $I \leq A(S)$ (cf. 3), and the sign of equality holds if and only if the given representation of S is eAC in Q^0.*

7. THEOREM. *Given S as in 1, suppose that $A(S) < +\infty$. Let $R_1, \ldots, R_m$ be a finite system of simply connected Jordan regions, without common interior points, whose sum is Q. Let $C_1, \ldots, C_m$ be the boundary curves of $R_1, \ldots, R_m$, and let $C_1^*, \ldots, C_m^*$ be the point-sets, in xyz space, that correspond to $C_1, \ldots, C_m$ by means of the given representation of S. Finally, let S_i be the surface determined by the given representation of S if (u, v) is restricted to R_i, $i = 1, 2, \ldots, m$. If the projections, upon the coordinate planes, of $C_1^*, \ldots, C_m^*$ are all of (planar) measure zero, then $A(S) = (AS_1) + \ldots + A(S_m)$.*

8. THEOREM. *Given S as in 1, we have always $a(S) \leq A(S)$. If either $A(S) < +\infty$ or $a(S) = 0$, then $a(S) = A(S)$. If S admits of a represen-*

tation of the form $z = f(x, y)$, where $f(x, y)$ is single-valued and continuous in a simply connected Jordan region, then always $a(S) = A(S)$.

9. THEOREM. *Given S as in 1, let p denote a plane through the origin in xyz space. Let ξ, η denote Cartesian coordinates in p and let $\kappa_p(\xi, \eta, Q^p)$ denote the essential multiplicity function associated with the orthogonal projection of S upon p. Let us define α_p as the integral of $\kappa_p(\xi, \eta, Q^p)$ if this integral exists, and let us put $\alpha_p = +\infty$ otherwise. Let P be a point on the unit sphere $x^2 + y^2 + z^2 = 1$, and let us put $\alpha(P) = \alpha_p$, where p is the plane through the origin that is perpendicular to the radius of the unit sphere that joins P to the origin. We have then the following generalization of a well-known theorem of Cauchy.*

If $A(S) < +\infty$, then $\alpha(P)$ is summable on the surface of the unit sphere, and its integral mean value (taken over the surface of the unit sphere) is equal to $A(S)/2$.

10. As regards the proofs, the decisive points may be indicated as follows. Due to the upward revision mentioned above, the lower area $a(S)$ can be shown to be equal to $A(S)$ in the cases described in the theorem in section 8. On the other hand, it can be shown that in spite of the upward revision the lower area $a(S)$ remains sufficiently close to previously studied lower areas to enable one to extend, after proper modifications, previously developed methods[1, 3, 4, 6] to $a(S)$. A few further remarks should be made concerning the theorem in 8. The second and third alternatives considered there are rather immediate consequences of previous results.[5, 3, 6] The proof of the fact that $a(S) = A(S)$ if $A(S) < +\infty$ is based essentially upon the following results. The given representation of S, interpreted as a transformation from Q into xyz space, gives rise to a monotone-light factorization, to a corresponding middle-space M, and finally to partial mappings corresponding to the proper cyclic elements of M.[1, 2, 8] Let S_1, S_2, ... be the surfaces determined by these partial mappings (if M is a dendrite, then these surfaces are missing, and the summations in the next statement are understood to be equal to zero). Then we have the formulas $A(S) = A(S_1) + A(S_2) + \ldots$, $a(S) = a(S_1) + a(S_2) + \ldots$. To secure the second one of these formulas, the upward revision of the lower area, mentioned above, seems to be indispensable. The first of these formulas was first studied by Morrey.[1, 2] His proof was based on a characterization theorem for representations defining the same surface S. Youngs found that this characterization theorem was false, derived an improved characterization theorem for the case of surfaces of the type of the 2-sphere, and proved[8] (for the 2-sphere case) the formula $A(S) = A(S_1) + A(S_2) + \ldots$. It seems, however, that even this improved characterization theorem of Youngs fails to hold for surfaces of the type of the 2-cell. However, the decisive formula $A(S) = A(S_1) + A(S_2) + \ldots$ can be established independently of the difficult topological issues suggested by the preceding observations.

11. It is entirely conceivable that the lower area $a(S)$, used in this study, is always equal to the essential area.[6] A more fundamental issue arises in connection with the theorem in 8. In view of that theorem, one may be excused for surmising that always $a(S) = A(S)$. If true, this fact would have far-reaching applications.

[1] Morrey, C. B., "An Analytic Characterization of Surfaces of Finite Lebesgue Area," *Am. Jour. Math.*, **57**, 692–702 (1935); **58**, 313–322 (1936).

[2] Morrey, C. B., "The Topology of (Path) Surfaces," *ibid.*, **57**, 17–50 (1935).

[3] Rado, T., and Reichelderfer, P., "A Theory of Absolutely Continuous Transformations in the Plane," *Trans. Am. Math. Soc.*, **49**, 258–307 (1941).

[4] Rado, T., and Reichelderfer, P., "Convergence in Length and Convergence in Area," *Duke Math. Jour.*, **9**, 527–565 (1942).

[5] Rado, T., "On Continuous Path-Surfaces of Zero Area," *Ann. Math.*, **44**, 173–191 (1943).

[6] Reichelderfer, P., "On Bounded Variation and Absolute Continuity for Parametric Representations of Continuous Surfaces," *Trans. Am. Math. Soc.*, **53**, 251–291 (1943).

[7] Youngs, J. W. T., "Curves and Surfaces," *Am. Math. Monthly*, **51**, 1–11 (1944).

[8] Youngs, J. W. T., "The Topological Theory of Fréchet Surfaces," *Ann. Math.*, **45** 753–785 (1944).

PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 31 April 15, 1945 Number 4

*VITAMIN A IN RELATION TO AGING AND TO LENGTH OF LIFE**

BY H. C. SHERMAN, H. L. CAMPBELL, MADELINE UDILJAK AND HELEN YARMOLINSKY

DEPARTMENT OF CHEMISTRY, COLUMBIA UNIVERSITY

Communicated February 21, 1945

Previous papers from this Department have shown (*a*) that a food supply which supports normal nutrition, generation after generation indefinitely, may still be capable of improvement with resulting advance in nutritional well-being,[1,2] and (*b*) that such a nutritional improvement of the norm may be due to enrichment of the initial dietary in one or more of its chemical factors, among which calcium, riboflavin and vitamin A may each play a major part.[3]

In the case of enrichment of the food with vitamin A, not only does the life history as a whole show improvement but in one series of experiments there was indication of a somewhat specific benefit to the life processes subsequent to the attainment of full maturity.[4]

The present paper records the results of further experimental studies of the effects of increased nutritional intake of vitamin A in the postponement of senility and the extension of the life span. In these experiments, as in the preceding series,[3,4] the basal diet (Laboratory No. 16) was a mixture of five-sixths ground whole wheat and one-sixth dried whole milk with table salt and distilled water (sometimes referred to as our Diet A). The degree of over-all adequacy of this diet for the nutrition of the experimental animals used (laboratory-bred albino rats of the Osborne-Mendel strain) is best illustrated by the fact that families of these rats are still thriving in our laboratory in the 58th generation on this diet. Its vitamin A value has not shown significant variations and averages approximately 3 International Units per gram, or 0.8 I. U. per calorie. Thus this level represents a vitamin A intake which has been shown to be *adequate* in the usual sense of the term but which as both our present and our previous experiments show, is not fully *optimal* inasmuch as enrichment of this diet in vitamin A results in better average life histories.

In the present experiments this level of 3 I. U. of vitamin A value per gram is compared with 6 I. U. and 12 I. U. in diets otherwise identical

(Laboratory Nos. 16, 360 and 361, respectively). The outstanding differences found both here and in our previous studies of the 3 I. U. and 6 I. U. levels are that the doubling of the already adequate level (of 3 I.U. per gram of air-dry food or 0.8 I. U. per food calorie) results in longer life for both sexes and a fully proportionate prolongation of the reproductive period in the females. As individual differences are relatively large, yet the trends have been consistent in the three series of experiments of this laboratory (Batchelder, 1934; Sherman and Campbell, 1937, and the present) we give in table 1 the weighted average results of the three series, i.e., of all the evidence of this kind available to date.

It is apparent from the evidence thus summarized in table 1 that, starting with a diet already adequate in the usual sense of the word, a doubling of the vitamin A value of the diet deferred old age and increased the length of life. When the vitamin A value was again doubled there was apparently still further benefit though the numbers of individuals at the highest level is not large enough to be regarded as entirely conclusive. It would be well to test in a similar way larger numbers at this 12 I. U. level with parallel cases at a level of 24 I. U. per gram.

TABLE 1

INFLUENCE OF THE VITAMIN A VALUE OF THE FOOD: EXPERIMENTS WITH RATS, NUMBER OF INDIVIDUALS AVERAGED IN EACH CASE ( )

	ON DIET WITH 3 I. U./G.	ON DIET WITH 6 I. U./G.	ON DIET WITH 12 I. U./G.
Reproductive period of females	(163) 265 days	(164) 312 days	(36) 369 days
Length of life:			
Of females	(163) 724 days	(165) 801 days	(36) 830 days
Of males	(112) 652 days	(108) 685 days	(24) 723 days

Without assuming either that the nutritional need for vitamin A is proportionate to the energy need or that the relative magnitudes of the two nutritional needs run parallel for rats and men, the experiments here reported may yet throw some light upon the problem of optimal allowances for human nutrition. The basal dietary of the present experiments with its 0.8 I. U. per calorie would correspond to 2400 I. U. for a man consuming 3000 calories a day. The doubled allowance, corresponding to 4800 I. U. per day, nearly all in the form of the vitamin itself as distinguished from provitamin, corresponds to the Recommended Allowances of the National Research Council. The experiments here reported, however, show that a further increase results in a greater benefit. This suggests that an allowance somewhat higher than that of the National Research Council might be more nearly optimal. This appears the more probable in view of the fact that most people probably live less protected lives than those of these experimental animals. As the benefit of a liberal intake is doubt-

less largely due to the laying-up of a reserve store in the body against emergencies which may increase the rate of destruction of the vitamin, the relations of intake to bodily storage are being studied quantitatively.

* Aided by grants from The Nutrition Foundation, Inc.

[1] Sherman, H. C., and Campbell, H. L., *Jour. Biol. Chem.*, **60**, 5 (1924).

[2] Sherman, H. C., and Campbell, H. L., *Proc. Nat. Acad. Sci.*, **14**, 852 (1928); *Jour. Nutrition*, **2**, 415 (1930).

[3] Sherman, H. C., and Campbell, H. L., *Jour. Nutrition*, **14**, 609 (1937).

[4] Sherman, H. C., Campbell, H. L., and Lanford, C. S., *Proc. Nat. Acad. Sci.*, **25**, 16 (1939).

THE LAW OF MASS ACTION IN EPIDEMIOLOGY, II

BY EDWIN B. WILSON AND JANE WORCESTER

HARVARD SCHOOL OF PUBLIC HEALTH

Communicated February 19, 1945

If one takes the equation for an epidemic when there are new susceptibles (recruits) coming into the population in the form[1]

$$\frac{dS}{dt} - A = \left(\frac{S}{m}\right)^p \left(\left.\frac{dS}{dt}\right|_{t-\tau} - A\right) \tag{1}$$

and introduces $u = \log (C/A)$, $C = Ae^u$, measuring case rates relative to the recruit rate A instead of relative to m, one finds as the equation for u

$$\left.\frac{du}{dt}\right|_t = \frac{pA}{m} e^{\frac{u(t-\tau)-u(t)}{p}} (1 - e^{u(t)}) + \left.\frac{du}{dt}\right|_{t-\tau}. \tag{2}$$

If t be advanced to $t + \tau/2$, one has

$$\left.\frac{du}{dt}\right|_{t+\tau/2} - \left.\frac{du}{dt}\right|_{t-\tau/2} = \frac{pA}{m} e^{\frac{u(t-\tau/2)-u(t+\tau/2)}{p}} (1 - e^{u(t+\tau/2)}). \tag{3}$$

The first approximation to this equation, neglecting first and higher derivatives upon the right, is

$$\frac{d^2u}{dt^2} = \frac{pA}{m\tau} (1 - e^u) \tag{4}$$

and corresponds to the approximation for the case $p = 1$ made by Soper in his discussion of periodicity. Equation (3) is indeed identical with his, except that $pA/m\tau$ replaces $A / m \tau$. The period of an infinitesimal oscillation is therefore

$$P = 2\pi\sqrt{\frac{m\tau}{pA}}. \tag{5}$$

The oscillations in epidemics of measles are not small. We have shown that, for approximations of this order in the absence of recruits ($A = 0$), we have

$$\frac{m}{p} = \frac{(\text{total cases})^2}{8(\text{peak cases})}, \tag{6}$$

where peak cases means the number of cases during one incubation period at the peak rate, and that if we consider higher approximations the value of m/p is not modified by more than perhaps 2 per cent unless the epidemic is very sharp (i.e., unless peak cases exceed one-sixth the total cases). We propose here to consider the modification in this relationship (6) that is due to the steady accession of susceptibles.

If we integrate (4) and determine the constant of integration so that $u = u_0$ when $du/dt = 0$, i.e., at the peak, we have

$$\frac{du}{dt} = \pm\sqrt{\frac{2pA}{m\tau}}\sqrt{u - e^u - u_0 + e^{u_0}}. \tag{7}$$

If u_1 be the negative root of $u - e^u - u_0 + e^{u_0}$, the half period will be obtained by integrating dt from $u = u_0$ to $u = u_1$ and the whole period will be found by doubling this result to be

$$P = \sqrt{\frac{m\tau}{2pA}}\left[2\int_{u_1}^{u_0}\frac{du}{\sqrt{u - e^u - u_0 + e^{u_0}}}\right] = \sqrt{\frac{m\tau}{pA}}\, f \tag{8}$$

where

$$f = \sqrt{2}\int_{u_1}^{u_0}\frac{du}{\sqrt{u - e^u - u_0 + e^{u_0}}} \tag{9}$$

and can be computed for different values of u_0 or $e^{u_0} = C_0/A$.

We now turn to finding the modification of (6) due to accession of recruits. If P be the period, the number of recruits during that time is PA and this must in turn be equal to the total cases, for in the hypothetical case under consideration everything must return to the same condition after one period. Hence

$$\text{Total cases} = PA = \sqrt{\frac{mA\tau}{p}}\, f. \tag{10}$$

As peak cases are $Ae^{u_0}\tau$, we have in lieu of (6)

$$\frac{m}{p} = \frac{e^{u_0}}{f^2} \frac{(\text{total cases})^2}{\text{peak cases}}. \tag{11}$$

In actual calculation we may obtain P from the total cases as P = total cases/A rather than from (8) and we may obtain m/p as

$$\frac{m}{p} = \frac{(\text{total cases})^2}{A\tau f^2}. \tag{12}$$

Here $A\tau$ are the recruits during an incubation period; to find f one must have a table of values in terms of u_0 or of $e^{u_0} = C_0/A$ computed from (9), viz.,

e^{u_0} = 1	2	4	6	10	15	20	30
$f = 2\pi$	6.4	7.1	8.0	9.7	11.5	13.1	15.3

The crucial test of any theory comes with the comparison of the theoretical results with the observed facts. If we had reliable values of m, of total cases, of recruits and of the peak case rate we could now determine the value of p appropriate to the particular relationship (12) derived from the theory. Fortunately, Hedrich[2] published a careful study of measles in Baltimore in which he has estimated the actual number of cases during each month from January, 1900, to December, 1931, and the number of intacts at the beginning of each month. By definition his intacts are children under 15 years who have not had measles. This should be close to the number of susceptibles S. In so far as the fundamental equation (1) is true, we should expect to find m as the value of S at the time when the case rate C was equal to the value one incubation period earlier. There are two such occasions each year around the time when the case rate is maximum and around the time when it is minimum. At minimum, cases are few and no great accuracy can be assigned to an estimate of the time when $C(t) = C(t - \tau)$; at maximum the cases are numerous in epidemic years but in non-epidemic years may be few and irregular. We have made the best estimates we can of the value of m at the peaks and in the troughs and find that the average value for the peaks is 67,000 and for the troughs 66,000. The average date of the peak (which varies from November to June) is estimated as around April 22 and the average date of the trough (which varies from August to November) is estimated as around September 22. The value m = 66,500 represents about $5^1/_2$ years of the estimates of recruits given by Hedrich.

If we take six of the most clear-cut epidemics, i.e., those that rise from low values of the case rate and die away to low values within a single epidemic year from September to August we find

Year	02–03	04–05	12–13	25–26	27–28	30–31
Cases	27,194	16,717	23,069	31,683	28,657	34,978
A (yr.)	11,100	11,000	11,700	12,000	11,500	10,500
P (yr.)	2.4	1.5	2.0	2.6	2.5	3.3
e^{μ_0}	10.0	5.7	7.8	11.4	8.8	12.4
m	58,632	59,817	61,146	70,629	65,562	64,054
p	3.4	6.1	4.3	3.7	3.2	2.6

These values of p are certainly not in the neighborhood of 1, and furthermore they show a great variation from epidemic to epidemic.

If we take other clear-cut epidemics of measles from a variety of places, we have no published estimate of m to use and no estimate of the true number of cases of measles. However, we may for some places find a record of measles for a long period of years over which both the child population and the number of cases of measles seem to show little or no trend, and on the reasonable assumption that from 90 to 100 per cent of all persons have measles before the age of 15 we may estimate the fraction φ of cases that are reported. We may also assume[3] that the value of m is 5.5 times the average annual population A under 15, i.e., $m = 5.5A$. Under these assumptions we have[4]

$$P = \frac{\text{total cases}}{A\varphi}, \qquad p = 0.23\left(\frac{A\varphi f}{\text{total cases}}\right), \qquad e^{\mu_0} = \frac{\text{peak cases}}{\varphi A/24}.$$

The results are given in table 1.

The value of P represents the number of years of recruits which are used up in the epidemic and must vary inversely with the estimate φ of the fraction of cases reported. As this estimate has been made by comparing the average cases reported with the recruits, it has been assumed that the reporting was equally good in all years. Such a period as that of 6.5 found for Minneapolis is not readily reconciled with the history of measles in that city before and after the great epidemic, namely,

30–31	31–32	32–33	33–34	34–35	35–36	36–37	37–38	38–39
1776	232	8148	276	18,022	3372	90	2677	4791

With the estimate of 6776 for annual recruits the total of 39,384 in 9 years would give a ratio of 65 per cent for reporting instead of the 41 per cent obtained from a longer run of years. Probably 65 per cent is high because the years 29–30 and 39–40 were very low. Measles is a very variable disease and any estimate based on a limited number of years must be subject to considerable error; furthermore there is no assurance that the fraction of reporting is the same from year to year, it may be higher in the years of large epidemics than in years relatively free from measles, or inversely. As may be seen from the formulae used in the calculation the value of P varies inversely with the fraction of reporting and that of p

TABLE 1

EPIDEMICS OF MEASLES WITH TOTAL CASES REPORTED, AND ESTIMATES OF RECRUITS PER ANNUM, AN ESTIMATE OF PEAK CASES, THE RATIO OF PEAK TO TOTAL CASES, AN ESTIMATE OF THE FRACTION REPORTED, THE VALUE OF $e^{M\theta}$ AND OF THE PERIOD P AND EXPONENT p

PLACE	YEAR	SEPT.	OCT.	NOV.	DEC.	JAN.	FEB.	MARCH	APRIL	MAY	JUNE	JULY	AUG.	TOTAL	A	φ	PEAK	PEAK[a]/TOTAL	$e^{M\theta}$	P	p
Albany, N. Y.	32–33	17	26	51	292	656	632	404	97	104	43	1	2	2,325	1713	0.45	351	0.15	10.9	3.0	2.5
Albany, N. Y.	35–36	1	5	26	111	282	701	547	252	149	108	54	0	2,236	1713	0.45	379	0.17	11.8	2.9	2.9
Albany, N. Y.	38–39	7	4	6	5	14	31	460	841	1049	309	24	7	2,757	1713	0.45	545	0.20	17.0	3.6	2.6
Berkeley, Calif.	30–31	8	16	2	4	18	88	362	661	372	95	15	2	1,643	1006	0.66	344	0.21	12.4	2.5	4.2
Berkeley, Calif.	33–34	6	2	2	18	66	104	153	344	334	90	7	8	1,134	1006	0.66	186	0.16	6.7	1.7	5.4
Berkeley, Calif.	38–39	2	5	14	64	377	865	1,260	372	92	13	6	3	3,073	1006	0.66	649	0.21	23.4	4.6	2.1
Binghamton, N. Y.	34–35	0	0	2	7	8	8	13	206	448	456	107	14	1,269	1161	0.49	246	0.19	10.4	2.2	4.4
Binghamton, N. Y.	37–38	3	0	1	4	4	15	122	269	296	141	25	0	880	1161	0.49	158	0.18	6.7	1.5	6.6
Binghamton, N. Y.	41–42	4	1	5	36	188	372	501	198	38	8	4	1	1,356	1161	0.49	256	0.19	10.8	2.4	4.0
Milwaukee, Wis.	31–32	11	9	11	26	99	587	1,892	4453	5544	1666	183	16	14,497	8927	0.62	2822	0.19	12.2	2.6	3.7
Milwaukee, Wis.	37–38	51	52	145	474	2534	8894	13,837	1595	221	75	31	13	27,922	8927	0.62	7266	0.26	31.5	5.0	2.2
Minneapolis, Minn.	34–35	19	74	234	1293	3828	6402	4,404	1347	310	67	38	6	18,022	6776	0.41	3449	0.19	29.8	6.5	1.3
New Haven, Conn.	30–31	1	5	39	42	108	350	1,428	1432	733	294	22	0	4,454	2587	0.51	767	0.17	13.9	3.4	2.5
New Haven, Conn.	34–35	2	2	7	43	104	301	925	2700	1428	326	20	4	5,862	2587	0.51	1420	0.24	25.8	4.4	2.4
New Haven, Conn.	38–39	3	11	3	12	49	170	479	1162	1480	698	101	20	4,188	2587	0.51	775	0.19	14.1	3.2	2.9
Niagara Falls, N. Y.	35–36	5	70	235	515	597	291	157	49	4	8	6	1	1,938	1352	0.41	307	0.16	13.3	3.5	2.2
Niagara Falls, N. Y.	37–38	2	0	4	3	15	126	329	398	251	73	35	5	1,241	1352	0.41	205	0.17	8.9	2.2	3.9
Niagara Falls, N. Y.	40–41	0	2	2	3	4	6	11	110	676	395	62	7	1,278	1352	0.41	362	0.28	15.7	2.3	5.9
Providence, R. I.	34–35	8	5	1	7	13	57	343	1351	1953	1279	241	17	5,275	4031	0.55	1003	0.19	10.9	2.4	4.0
Providence, R. I.	36–37	0	0	9	77	422	811	1,182	711	472	129	31	4	3,848	4031	0.55	589	0.15	6.4	1.7	5.1
Syracuse, N. Y.	31–32	0	1	1	3	83	728	2,323	2058	1171	587	191	4	7,150	3106	0.52	1156	0.16	17.2	4.4	1.7
Syracuse, N. Y.	38–39	4	0	10	3	89	286	405	697	1224	497	436	7	3,658	3106	0.52	640	0.17	9.5	2.3	4.0
Syracuse, N. Y.	41–42	5	0	0	2	2	25	107	324	1094	2255	610	74	4,498	3106	0.52	1191	0.26	17.7	2.8	4.5

[a] We have stated that ordinarily Peak/Total $<1/6$ and that in that case our approximate formulae should be good to within 2 per cent. For many of the epidemics here listed the ratio is greater than $1/6$ and consequently the approximation will not be so good; we desired, however, to have a wide range of epidemics, regardless of the accuracy of the formulae, particularly as the difficulties in estimating the necessary quantities are such as to introduce a liability to a considerable error in the values of P and p. It should be observed that the values of p for these 23 epidemics are not correlated appreciably with the ratio Peak/Total.

varies directly with the square of that fraction multiplied by f. Thus should we use 65 per cent in place of 41 per cent, we should get $P = 4.1$ and $p = 2.2$ in place of $P = 6.5$ and $p = 1.3$.

As a result of the difficulty of making reliable estimates of the factors of reporting the values of P and p which are entered in table 1 cannot be regarded as individually well determined; but the conclusion from all of them and from the results obtained for Baltimore seems inescapable that: In so far as the relationship $C = (S/m)^p C_{-1}$ may be valid for the course of epidemics of measles, the value of p which is appropriate to the relationship between peak cases and total cases cannot be considered to be unity, i.e., the simple law of mass action which has been so widely used is not in accord with the facts, and it is doubtful whether any fixed value can be brought into satisfactory accord with the course of epidemics in different years and in different places.[5]

[1] The notation will be essentially that of earlier papers, see these PROCEEDINGS, **31**, 24–34 (1945).

[2] Hedrich, A. W., "Monthly Estimates of the Child Population 'Susceptible' to Measles 1900–1931, Baltimore, Md.," *Amer. Jour. Hyg.*, **17**, 613–636 (1933). Such estimates might appear easy to make. If one starts with any level of susceptibles at the beginning of any month, adds the number of recruits and subtracts the number of cases, one obtains the number of susceptibles at the beginning of the next month. The difficulty of scaling up the number of cases reported to the true number of cases is however so serious that the process is not easy to carry out over a long period of time without getting unreasonably high or unreasonably low values of the susceptibles at some times. Hedrich's work seems to be carefully and critically done and we shall base our calculations upon his figures.

[3] The age distribution of reported cases of measles for New York State (exclusive of New York City) and for Massachusetts for the years specified is

NEW YORK STATE (EXCLUSIVE OF N. Y. CITY)

	1932		1933		1934		1935		1936		
AGE	CASES	%	CASES	%	CASES	%	CASES	%	CASES	%	MEAN, %
Under 1	927	1.87	703	1.99	577	2.01	710	1.78	537	1.94	1.92
1	2,015	4.06	1,463	4.15	1,098	3.82	1,339	3.36	1,168	4.21	3.92
2	2,714	5.47	1,879	5.33	1,435	5.00	1,768	4.43	1,422	5.13	5.07
3	3,182	6.41	2,150	6.10	1,658	5.77	2,184	5.47	1,669	6.02	5.95
4	3,634	7.32	2,516	7.13	1,970	6.86	2,442	6.12	1,910	6.89	6.86
5	5,401	10.88	3,806	10.79	3,279	11.41	3,825	9.59	3,171	11.43	10.82
6	7,093	14.29	5,220	14.80	4,434	15.44	5,154	12.92	4,407	15.89	14.67
7	6,263	12.62	4,342	12.31	3,817	13.29	4,865	12.19	3,875	13.97	12.88
8	5,118	10.31	3,427	9.72	2,916	10.15	3,673	9.20	2,903	10.47	9.97
9	3,245	6.54	2,290	6.49	1,990	6.93	2,468	6.18	1,777	6.41	6.51
10–14	6,759	13.62	4,945	14.02	4,007	13.95	6,796	17.03	3,468	12.50	14.22
15–19	1,765	3.56	1,328	3.75	833	2.90	2,340	5.86	765	2.76	3.77
20+	1,505	3.03	1,204	3.41	712	2.48	2,340	5.86	666	2.40	3.44
	49,620	99.98	35,268	99.99	28,726	100.01	39,904	99.99	27,738	100.02	100.00

The variation from year to year is notable. The large percentage of older cases in 1935 is especially noteworthy; the percentages at all ages under 10 are below the average

of the percentages and are above the averages at all ages over 10. Contrariwise in 1936 the percentages are above the averages up to age 9 and below them after that age.

MASSACHUSETTS

	1936		1937		1938		1939		1940		
AGE	CASES	%	CASES	%	CASES	%	CASES	%	CASES	%	MEAN, %
Under 1	527	1.95	389	1.91	334	3.27	422	1.62	619	2.92	2.33
1	1,181	4.38	896	4.40	571	5.60	1,062	4.08	1,071	5.06	4.70
2	1,498	5.55	1,293	6.35	651	6.38	1,500	5.76	1,467	6.93	6.19
3	1,864	6.91	1,426	7.00	781	7.66	1,841	7.07	1,924	9.08	7.54
4	2,431	9.01	1,792	8.80	1,025	10.05	2,079	7.98	1,828	8.63	8.89
5	3,403	12.61	2,260	11.10	1,339	13.13	2,954	11.34	2,448	11.56	11.95
6	4,977	18.44	3,749	18.41	1,950	19.12	4,437	17.03	3,500	16.53	17.91
7	4,089	15.15	3,068	15.06	1,406	13.78	3,818	14.66	3,075	14.52	14.63
8	2,445	9.06	2,156	10.59	794	7.78	2,813	10.80	1,983	9.36	9.52
9	1,249	4.63	1,058	5.19	371	3.64	1,707	6.55	1,185	5.60	5.12
10–14	2,470	9.15	1,677	8.23	609	5.97	2,592	9.95	1,378	6.51	7.96
15–19	443	1.64	322	1.58	212	2.08	475	1.82	371	1.75	1.77
20+	410	1.52	280	1.37	157	1.54	347	1.33	330	1.56	1.46
	26,987	100.00	20,366	99.99	10,200	100.00	26,047	99.99	21,179	100.01	99.97

Again the variation from year to year is large. The percentages in 1939 are below the averages for all ages under 7. In comparison with the mean percentages of New York State the Massachusetts averages are higher under 8 and lower over 8; the cumulated percentages under 8 are Massachusetts 74.2, New York 62.1. It is clear that no percentage distribution can be assigned that is valid in both States in all years. How much differential there is in the factors of reporting by age in either State is unknown. The effect of the different age distributions in the two States could be eliminated but would make no really substantial modification.

If one bases an actuarial calculation upon the mean percentages one finds that the average number of immunes in the population is more than $5^1/_2$ years of recruits in Massachusetts and still more in New York State. On the other hand the figures given by Selwyn Collins (Public Health Reports, April 5, 1929) obtained from a large number of surveys in which was tabulated by age the percentage of children who had had measles, indicate much higher attack rates for children at early ages than those found here in either State. His fitted curve

$$y = 89\ (1 - e^{0.00886-0.04863x-0.02899x^2})$$

for the percentage who have had measles by age x, while representing well the observations has the obvious defect which inheres in all such series, namely, that the asymptotic percentage is too far below 100 to be representative of the true situation with respect to measles. There is no telling how the percentages should be scaled up to represent the true situation but the figures obtained from the curve give, respectively, 10.5, 12.7, 13.3, 12.3, 10.5, at ages 1, 2, 3, 4, 5 in place of 4.7, 6.2, 7.5, 8.9, 12.0 in Massachusetts. Conversion of the Massachusetts figures to rates would modify the percentages in a minor way. Clearly any estimate of the average number of susceptibles in the population based on rates derived from Collins' distribution, however, those rates were altered to come more nearly to representing the true situation with respect to immunity in the population, would be well below that derivable from the reported cases.

The evidence is as a whole indicative of a value for m in the neighborhood of $5^1/_2$ years of recruits.

[4] Peak cases as reported may be estimated by inspection as slightly more than half the cases in the highest month. The difference between 30 and 31 day months is small

enough to be disregarded in view of statistical fluctuations and the inherent inaccuracies of estimating the factor of reporting. The short month February may better be adjusted by taking from January and March an allowance for the last day of January and the first day of March, leaving these as 30 day months, in ordinary years, with slightly different allowances in leap years. If it be assumed that the case rate is parabolic for the three highest months and that the cases are, respectively, k_{-1}, k_0, k_1 in sequence, then on the assumption that τ is half a month

$$C_0\tau = \text{"peak cases"} = \frac{26k_0 - k_{-1} - k_1}{48} + \frac{(k_1 - k_{-1})^2}{16(2k_0 - k_{-1} - k_1)}.$$

As the theoretical solution in the absence of recruits ($A = 0$) for the epidemic curve is the sech² curve, and as the effect of the recruits is probably small in the three highest months of the epidemic, and finally as the probability and sech² curves have been used more for fitting observed cases than the parabola, it might be better to estimate peak cases by fitting a sech² curve instead of a parabola; we have indeed used the sech² curve in a number of cases but have come to the conclusion that the extra work involved is not justified by the slight increase in accuracy which may thereby be obtained.

[5] It is interesting to make some calculations for the data Soper gave for Glasgow for the years 1901–1916 which he seemed to think were of the forty years with which he worked those best suited to test his theory. One noticeable difference between his data and Hedrich's for Baltimore or that for other American cities is the infrequency with which clear-cut epidemics, which rise from few cases in one summer and die away to few cases in the next, are found in Glasgow. For five epidemic years we find, however,

YEAR	TOTAL CASES	PEAK	A	φ	P	$c\omega^2$	f	p
03–04	13,544	1177	800	0.55	1.2	2.7	6.6	4.8
07–08	21,484	1988	800	0.55	1.9	4.5	7.3	2.3
09–10	22,533	3153	800	0.55	2.0	7.2	8.5	2.9
11–12	18,176	2158	800	0.55	1.6	4.9	7.5	3.4

In making the calculations we have taken Soper's estimate of births as 25,500 and reduced it by about 20 per cent to allow for deaths at early ages, leaving recruits as $A = 800$ per fortnight. We have taken φ as 0.55 because Soper estimated the reporting as a little less than 50 per cent of births. Further we have taken m as 4 years of recruits in place of $5^1/_2$ years because he mentions that the average age of measles in one epidemic was $4^1/_2$ years. The formula for p then becomes $104(440\ f/\text{total cases})^2$. The average value of P is 1.7 years which corresponds fairly well to his figure of a little under two years, but the values of p are certainly not reconcilable with the assumption $p = 1$. Soper appeared to take m as 2 years of births (recruits?) and this would cut our values of p in two. He did, however, note that m would appear to be 3.87 years of births (something probably in excess of four years of recruits) judged from the average age of cases but rejected this estimate for reasons that seem to us of doubtful cogency.

AXIOMATIC APPROACH TO HOMOLOGY THEORY

BY SAMUEL EILENBERG AND NORMAN E. STEENROD

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MICHIGAN

Communicated February 21, 1945

1. *Introduction.*—The present paper provides a brief outline of an axiomatic approach to the concept: homology group. It is intended that a full development should appear in book form.

The usual approach to homology theory is by way of the somewhat complicated idea of a complex. In order to arrive at a purely topological concept, the student of the subject is required to wade patiently through a large amount of analytic geometry. Many of the ideas used in the constructions, such as orientation, chain and algebraic boundary, seem artificial. The motivation for their use appears only in retrospect.

Since, in the case of homology groups, the definition by construction is so unwieldy, it is to be expected that an axiomatic approach or definition by properties should result in greater logical simplicity and in a broadened point of view. Naturally enough, the definition by construction is not eliminated by the axiomatic approach. It constitutes an existence proof or proof of consistency.

2. *Preliminaries.*—The concepts of a topological space and of a group are assumed to be known. The symbol (X, A) stands for a pair consisting of a topological space X and a closed subset A. A map $f:(X, A) \to (Y, B)$ of one such pair into another is a continuous map of X into Y which maps A into B. In case A is the vacuous set (X, A) is written as (X). If f_0, f_1 are two maps of (X, A) into (Y, B), they are homotopic if there exists a homotopy $f(x, t)$ connecting the two maps of X into Y such that $f(x, t) \in B$ for any $x \in A$ and all t.

3. *Basic Concepts.*—The fundamental concept to be axiomatized is a function $H_q(X, A)$ (called the *q-dimensional, relative homology group of X mod A*) defined for all triples consisting of an integer $q \geqq 0$ and a pair (X, A). The value of the function is an abelian group.

The first subsidiary concept is that of boundary. For each $q \geqq 1$ and each (X, A), there is a homomorphism

$$\partial: H_q(X, A) \to H_{q-1}(A)$$

called the *boundary operator.*

The second subsidiary concept is that of the induced homomorphism. If f is a map of (X, A) into (Y, B) and $q \geqq 0$, there is an attached homomorphism

$$f_*: H_q(X, A) \to H_q(Y, B)$$

called the *homomorphism induced by f*.

4. *Axioms.*—These three concepts have the following properties.

AXIOM 1. *If f = identity, then f_* = identity.*

That is to say, if f is the identity map of (X, A) on itself, then f_* is the identity map of $H_q(X, A)$ on itself.

AXIOM 2. $(gf)_* = g_*f_*$.

Explicitly, if $f:(X, A) \to (Y, B)$ and $g:(Y, B) \to (Z, C)$, then the combination of the induced homomorphisms $f_*: H_q(X, A) \to H_q(Y, B)$ and $g_*: H_q(Y, B) \to H_q(Z, C)$ is the induced homomorphism $(gf)_*: H_q(X, A) \to H_q(Z, C)$.

An immediate consequence of Axioms 1 and 2 is that homeomorphic pairs (X, A) and (Y, B) have isomorphic homology groups.

AXIOM 3. $\partial f_* = f_* \partial$.

Explicitly, if $f:(X, A) \to (Y, B)$ and $q \geqq 1$, the axiom demands that two homomorphisms of $H_q(X, A)$ into $H_{q-1}(B)$ shall coincide. The first is the combination of $\partial: H_q(X, A) \to H_{q-1}(A)$ followed by $(f|A)_*: H_{q-1}(A) \to H_{q-1}(B)$. The second is the combination of $f_*: H_q(X, A) \to H_q(Y, B)$ followed by $\partial: H_q(Y, B) \to H_{q-1}(B)$.

AXIOM 4. *If f is homotopic to g, then $f_* = g_*$.*

Definition: The *natural system* of the pair (X, A) is the sequence of groups and homomorphisms

$$\ldots \to H_q(X) \to H_q(X, A) \to H_{q-1}(A) \to H_{q-1}(X) \to \ldots \to H_0(X, A)$$

where $H_q(X) \to H_q(X, A)$ is induced by the identity map $(X) \to (X, A)$, $H_q(X, A) \to H_{q-1}(A)$ is the boundary operation, and $H_{q-1}(A) \to H_{q-1}(X)$ is induced by the identity map $(A) \to (X)$.

AXIOM 5. *In the natural system of (X, A) the last group, $H_0(X, A)$, is the image of $H_0(X)$. In any other group of the sequence, the image of the preceding group coincides with the kernel of the succeeding homomorphism.*

At first glance, this axiom may seem strange even to one familiar with homology theory. It is equivalent to three propositions usually stated as follows: (1) the boundary of a cycle of X mod A bounds in A if and only if the cycle is homologous mod A to a cycle of X; (2) a cycle of A is homologous to zero in X if and only if it is the boundary of a cycle of X mod A; (3) a cycle of X is homologous to a cycle of A if and only if it is homologous to zero mod A.

Definition: An open set U of X is strongly contained in A, written $U \Subset A$, if the closure $\bar{U}$ is contained in an open set $V \subset A$.

AXIOM 6. *If $U \Subset A$, then the identity map: $(X - U, A - U) \to (X, A)$ induces isomorphisms $H_q(X - U, A - U) \cong H_q(X, A)$ for each $q \geqq 0$.*

This axiom expresses the intuitive idea that $H_q(X, A)$ is pretty much independent of the internal structure of A.

AXIOM 7. *If P is a point, then $H_q(P) = 0$ for $q \geqq 1$.*

A particular reference point P_0 is selected, and $H_0(P_0)$ is called the *coefficient group* of the homology theory.

5. *Uniqueness.*—On the basis of these seven axioms, one can deduce the entire homology theory of a complex in the usual sense. Some highlights of the procedure are the following.

If σ is an n-simplex, and $\dot\sigma$ its point-set boundary, then $H_n(\sigma, \dot\sigma)$ is isomorphic to the coefficient group. Further, $H_q(\sigma, \dot\sigma) = 0$ for $q \neq n$, and the boundary operator $\partial : H_n(\sigma, \dot\sigma) \to H_{n-1}(\dot\sigma)$ is an isomorphism onto for $n > 1$, and into for $n = 1$.

Let f be the simplicial map of σ on itself which interchanges two vertices and leaves all others fixed. Then, for any $z \in H_n(\sigma, \dot\sigma)$, we have $f_*(z) = -z$. This permits the usual division of permutations into the classes of even and odd, and leads naturally to a definition of orientation—a concept which is quite troublesome in the usual approach.

Definition: Let H, H' be two homology theories satisfying Axioms 1 through 7. A *homomorphism*

$$h : H \to H'$$

is defined to be a system of homomorphisms

$$h(q, X, A) : H_q(X, A) \to H_q'(X, A)$$

defined for all q, (X, A), which commute properly with the boundary operator and induced homomorphisms:

$$h(q - 1, A)\partial = \partial' h(q, X, A), \qquad h(q, Y, B)f_* = f_*' h(q, X, A). \quad \text{(I)}$$

If h gives an isomorphism of the coefficient groups $h(0, P_0) : H_0(P_0) \cong H_0'(P_0)$, then h is called a *strong homomorphism.* If each $h(q, X, A)$ is an isomorphism, then h is called an *equivalence* and H and H' are called *equivalent.*

Since the usual homology theory of complexes is deducible from the axioms, there follows the

UNIQUENESS THEOREM: *Any two homology theories having the same coefficient group coincide on complexes.*

Explicitly, if $i : H_0(P_0) \cong H_0'(P_0)$ is an isomorphism between the coefficient groups of H and H', then isomorphisms

$$h(q, X, A) : H_q(X, A) \cong H_q'(X, A)$$

can be defined for X a complex, A a subcomplex such that $h(0, P_0)$ coincides with i, and the relations (I) hold in so far as they are defined (f need not be simplicial). Indeed, there is just one way of constructing $h(q, X, A)$. The uniqueness theorem implies that any strong homomorphism $h : H \to$

H' is an equivalence as far as complexes are concerned. In view of Axiom 4, the uniqueness theorem holds for spaces having the same homotopy type as complexes. These include the absolute neighborhood retracts.

6. *Existence.*—As is to be expected, homology theories exist which satisfy the axioms. Both the Čech homology theory H^1 and the singular homology theory H^0 satisfy the axioms. This is fairly well known, although the proofs of some of the axioms are only implicitly contained in the literature. It is well known that the two homology theories differ for some pairs (X, A). Thus, the axioms do not provide uniqueness for all spaces.

The surprising feature of H^0 and H^1 that appears in this development is that they play extreme roles in the family of all homology theories, and have parallel definitions. They can be defined as follows: The homology groups of the simplicial structure of a complex (using chains, etc.) are defined as usual. (As a first step of an existence proof, this is quite natural since the definition has been deduced from the axioms.) Using maps $K \rightarrow X$ of complexes into the space X, the singular homology groups $H^0{}_q(X, A)$ can be defined using a suitable limiting process. Similarly using maps $X \rightarrow K$ of the space into complexes, the Čech homology groups $H^1{}_q(X, A)$ are obtained. It is then established that H^0 and H^1 are minimal and maximal in the family of all homology theories with a prescribed coefficient group in the sense that, if H is any homology theory, there exist strong homomorphisms $H^0 \rightarrow H \rightarrow H^1$. This is an indication of how it is possible to characterize H^0 or H^1 by the addition of a suitable Axiom 8.

7. *Generalizations.*—A suitable refinement of the axioms will permit the introduction of topologized homology groups.

Cohomology can be axiomatized in the same way as homology. It is only necessary to reverse the directions of the operators ∂ and f_* in the above axioms and make such modifications in the statements as these reversals entail. The analogous uniqueness theorems can be proved.

The products of elements of two cohomology groups with values in a third (in the usual sense) may also be axiomatized and characterized uniquely.

AWARDS TO THE ACADEMY AND RESEARCH COUNCIL

The National Academy of Sciences was incorporated by an act of the Congress in 1863 during the Civil War to mobilize the best scientific talent of the country for service to the Government in war and in peace. The National Research Council was organized in 1916 at the request of the President by the National Academy of Sciences as a measure of national

This

Award of Distinction

is presented to

National Research Council

in recognition of its fundamental contributions to public health in the field of the medical sciences and also in recognition of its essential services to our country in World Wars I and II.

By

American Pharmaceutical Manufacturers' Association

in New York City

December 11, 1944

Ordnance

Distinguished Service Award

presented to

National Academy of Sciences

In recognition of outstanding and meritorious scientific services, in war and peace, for the development, manufacture and maintenance of Ordnance materiel

Award Authorized 20 July 1944

preparedness and was perpetuated in accord with an Executive Order of the President dated May 11, 1918. The Academy and the Research Council have been active since their establishment both in war and in peace, and particularly during the current war. These activities have been formally recognized by the presentation to the Academy at its Autumn Meeting, November 15, 1944, of the Ordnance Distinguished Service Award and by the presentation to the Research Council on December 11, 1944, of an Award of Distinction by the American Pharmaceutical Manufacturers' Association. Reproduction of the documents of award are published herewith by authority of the President of the Academy and of the Chairman of the Research Council.

IRREGULAR PROJECTIVE INVARIANTS

BY EDWARD KASNER AND JOHN DE CICCO

DEPARTMENTS OF MATHEMATICS, COLUMBIA UNIVERSITY AND ILLINOIS INSTITUTE OF TECHNOLOGY

Communicated March 16, 1945

1. We shall begin, in this paper, the classification of irregular analytic elements with respect to the eight-parameter group of collineations in the plane. Halphen showed that a regular analytic element has two relative projective invariants of fifth and seventh orders and an absolute differential invariant of the seventh order. We show that an irregular element with a simple cusp possesses two relative invariants of the seventh and eighth orders, and therefore an absolute invariant of the eighth order.

Kasner has given a complete classification of the irregular analytic elements according to the type of absolute invariant of the lowest order with respect to the infinite group of conformal transformations.[1] The authors have made a similar classification with respect to the infinite group of all analytic point transformations.[2]

2. The configuration that is to be considered is an analytic arc (regular or irregular) together with a specific point of the arc. This compound object may be termed an analytic element. If the specific point is taken as the origin, the most general analytic element may be defined by setting x and y equal to two power series in integral powers of a parameter t without constant terms. Let the integer $p \geqq 1$, be the minimum of the two exponents of the leading terms in the two power series of t; we may assume this to appear in the x-series. Therefore our analytic element may be written in the form

$$y = c_p x^{p/p} + c_q x^{q/p} + c_{q+1} x^{(q+1)/p} + \dots, \qquad (1)$$

where $q > p \geqq 1$ and $c_q \neq 0$. Of course, we may have $q = p + 1$.

If only integral powers of x appear in (1), then our element is called regular. Otherwise, it is said to be irregular.

For irregular elements, both p and q are arithmetical projective invariants. Therefore we define p as the *index* and q as the *rank*. All irregular elements obtained by taking arbitrary values of the coefficients but with fixed values of the integers p and q, we shall define as the single *species* (p, q).

3. Consider the species (2, 3). Any irregular element of this species may be written in the form

$$y = c_2x^{2/2} + c_3x^{3/2} + c_4x^{4/2} + c_5x^{5/2} + \ldots, \tag{2}$$

where $c_3 \neq 0$.

Under the projective group, we have proved that

$$\alpha = \frac{1}{c_3^{24}}\,[2c_3^3c_7 - 6c_3^2c_4c_6 - 3c_3^2c_5^2 + 12c_3c_4^2c_5 - 5c_4^4]^5, \tag{3}$$

is a relative differential invariant of weight four.

Thus the analytic element (2) possesses a relative differential invariant α of order seven. This is the lowest possible order.

Under the projective group, we have proved that

$$\beta = \frac{1}{c_3^6}\begin{bmatrix} 9c_3^4c_8 - 30c_3^3c_4c_7 - 36c_3^3c_5c_6 + 72c_3^2c_4^2c_6 + \\ 81c_3^2c_4c_5^2 - 244c_3c_4^3c_5 + 48c_4^5 \end{bmatrix} \tag{4}$$

is a relative differential invariant of weight one.

Our new relative invariant β is of order eight.

By (3) and (4), it follows that *the irregular analytic element* (2) *possesses the absolute differential projective invariant*

$$I_8 = \beta^4/\alpha. \tag{5}$$

This absolute invariant is of order eight, degree twenty and total subscript weight eighty.

Obviously, any analytic element has an infinite number of projective invariants.

Under the conformal group, Kasner proved that the species (2, 3) has no invariants, any such element being reducible to the canonical form $y = x^{3/2}$. Therefore this is also true under the infinite group of all regular analytic point transformations.

In our later work, we shall consider the projective differential invariants of other species of analytic elements (p, q). We may describe our invariants as *local*, since they apply in the neighborhood of the singularity of the curve which may be either algebraic or transcendental. If we allow correlations as well as collineations, we show that the *species* (p, q) *is equivalent to the species* $(q - p, q)$. This is a new duality theorem.[3]

[1] Kasner, "Conformal Classification of Analytic Arcs or Elements, Poincaré's Local Problem of Conformal Geometry," *Trans. Am. Math. Soc.*, **16**, 333–349 (1915). The theory of a pair of regular arcs, including the horn angle, is given in "Conformal Geometry," *Proceedings* Cambridge International Congress Mathematicians, 1912, and a paper appearing in *Scripta Mathematica*, 1945.

[2] Kasner and De Cicco, "The General Invariant Theory of Irregular Analytic Arcs or Elements," *Ibid.*, **51**, 232-254 (1942). Also in Publications of the Illinois Institute of Technology (1943.)

[3] See for regular curves Halhhen's dissertation 1878, and his collected works Vol. 2. Also Lane, "Projective Differential Geometry," Chicago Press, 1942.

EULER'S THREE BIQUADRATE PROBLEM

BY MORGAN WARD

CALIFORNIA INSTITUTE OF TECHNOLOGY

Communicated March 19, 1945

1. Euler's problem of whether the sum of three biquadrates can be a biquadrate; that is, whether the diophantine equation

$$x^4 + y^4 + z^4 = w^4 \tag{1}$$

has any (non-trivial) integer solutions, has never been solved.[1] The problem is a hard one; indeed, a modern investigator has stated: ". . . it would be difficult to mention any other [problem] which has yielded so little to the efforts of those who have attempted its solution."

The most that is known to date is that there is no solution of (1) with[3] $w < 1024$. I have recently proved that *there is no solution of* (1) *with*

$$w < 10{,}000. \tag{2}$$

This result makes it appear probable that there are no solutions of (1) whatever, especially since several closely allied soluble diophantine equations such as $x^4 + y^4 + z^4 + t^4 = w^4$, $x^4 + y^4 = w^4 + t^4$, $x^4 + 2y^4 + 2z^4 = w^4$ are known to have comparatively small solutions.[4]

2. The first step of the proof is to reduce the solution of (1) to the solution of another diophantine equation containing more variables but with the variables subjected to a number of restrictive conditions which it is unnecessary to state here:

$$u^4 + v^4 = 2\epsilon kl(e^8l^2 + 2^{18+8\sigma+2\epsilon}d^8k^2). \tag{3}$$

The old variables are easily expressed in terms of the new; for example,

$$w = 2^{4\sigma+9+\epsilon}d^4k + e^4l. \tag{4}$$

Equation (1) has a solution if and only if equation (3) has a solution with d, e, k, l, u and v positive integers. The exponent σ is a positive integer or zero, and the exponent ϵ is either zero or one.

The inequality (2) in conjunction with (4) immediately restricts σ, d, e, k and l to a finite number of choices; in fact,

$$\sigma \leqq 1, d \leqq 1, e \leqq 9, k \leqq 17 \text{ and } l \leqq 9488. \tag{5}$$

3. The second step of the proof is to discuss (3) for each of the cases given by (5). The most difficult case turns out to be when $\sigma = 0$, $\epsilon = 1$ and $d = e = k = 1$. (3) then becomes

$$u^4 + v^4 = 2l(l^2 + 1024^2) \tag{6}$$

with

(i) $l < 8976$.

The restrictions on the variables in (3) alluded to in Section 2 tell us that

(ii) Every prime factor of l and $l^2 + 1024^2$ is congruent to one modulo eight.

On considering (6) modulo 5 and modulo 13, we find that

(iii) $l \equiv 4 \pmod 5$,
(iv) $l \equiv 3, 4, 5, 7, 10$ or $12 \pmod{13}$.

The conditions (i)–(iv) reduce the possible choices of l to twenty-nine numbers: 289, 449, . . ., 8689.

The other cases lead to even fewer choices of l and the other variables in (5).

4. The third step of the proof is to dispose of the cases which survive after all conditions of the type (i)–(iv) just described have been applied. For example, in the case given by (6), we have to show by the composition formulae for products of sums of squares that $2l(l^2 + 1024^2)$ is not a sum of two biquadrates for twenty-nine numerical values of l. This last step is easily carried out, and the proof is complete.

5. The most laborious feature in the proof is the necessity for factoring several numbers greater than ten million, the extent of the present factor tables. For example, in the case discussed in Section 3, it is necessary to factor the number $l^2 + 1024^2$ not only in order that condition (ii) may be applied, but also in order to apply the final restrictions by composition of sums of squares. This work was performed with the aid of a calculating machine by the factor stencil method of D. N. Lehmer and J. D. Elder.[5] Whenever the stencils indicated that the number was a prime, the fact was confirmed by D. H. Lehmer's[6] method based on the converse of Fermat's theorem.

In order to insure accuracy, all the attendant numerical work and the algebra of determining the cases in step two was checked twice at different times. Complete details of the proof will appear elsewhere.

[1] In L. E. Dickson's *History of the Theory of Numbers*, vol. 2, p. 648—there is a statement that might lead one to infer that the impossibility of (1) was proved by A. Werebrusow (*L'Intérmediare des Mathématiciens*, **21**, 161 (1914)). A fatal lacuna in Werebrusow's proof was pointed out by E. T. Bell (*Mathematics Student*, **4**, 78 (1936)).

[2] Mordell, L. J., "The Present State of Some Problems in the Theory of Numbers," *Nature*, **121**, 138 (1928).

[3] Aubry, L., *Sphinx-Oedipe*, **7**, 45–46 (March, 1912).

[4] For example, we have Norrie's well-known result that

$$30^4 + 120^4 + 272^4 + 315^4 = 353^4.$$

[5] Lehmer, D. N., and Elder, J. D., "Factor Stencils," Carnegie Institution, Washington (1939).

[6] Lehmer, D. H., *Amer. Math. Monthly*, **43**, 347–354 (1936).

ERRATUM

In the article, "Dominance Modification and Physiological Effects of Genes," by L. C. Dunn and S. Gluecksohn-Schoenheimer, *Proc. Nat. Acad. Sci.*, **31**, 82 (1945), the formula in the middle of line 1, page 83, should read "138 Sd + ($\chi^2 = 5.26$, $p = 0.02$)" instead of "138 Sd + $X^{2=5.26}_{p=0.2}$."

PROCEEDINGS

OF THE

NATIONAL ACADEMY OF SCIENCES

Volume 31 May 15, 1945 Number 5

---

*STRAIN SPECIFICITY AND PRODUCTION OF ANTIBIOTIC SUBSTANCES. IV. VARIATIONS AMONG ACTINOMYCETES, WITH SPECIAL REFERENCE TO ACTINOMYCES GRISEUS**

By Albert Schatz and Selman A. Waksman

New Jersey Agricultural Experiment Station and Rutgers University

Communicated March 15, 1945

Introductory and Historical.—Every student of the actinomycetes has been impressed by the striking variations which these organisms undergo in culture. Some of these variations have long been recognized to be quantitative in nature, whereas others are also qualitative. These variations are found to occur in the pigment production by the culture, in the intensity of the pigment, as well as in its specific nature; also whether it is dissolved in the medium or retained in the mycelium. Further variations are found in the production of enzyme systems, especially proteases and diastases, and in other cultural and physiological characteristics. The formation of aerial mycelium and the manner of sporulation of the culture are among the most significant variable factors. These variation phenomena are further complicated by the fact that frequently sectors or saltants are produced which differ from the rest of the colony with respect to the formation of aerial mycelium or pigmentation; upon isolation, such sectors have given cultures that have often been considered as variants.

These variations have caused considerable confusion in the recognition of fixed species or types for the characterization and classification of this group of organisms. They have even led certain investigators[5, 8] to question the constancy of many actinomyces types as recognized by species descriptions. In illustrating this phenomenon, Waksman[8] directed attention, as far back as 1919, to the variability in the proteolytic mechanism of *Actinomyces griseus*, the organism that forms the subject of this paper. One of the more recent students of variations among actinomycetes[4] came to the conclusion that the morphological and physiological properties of these organisms, when cultured under the same environmental conditions, are constant. The changes observed were believed to be either transitory

modifications, or quantitative rather than qualitative variations, long-time changes being produced under the influence of altered environmental conditions. In addition to these, however, sudden changes of a permanent and hereditary nature were obtained without apparently being affected by the environment. These new forms or variants, which were believed to be rather rare among actinomycetes, differed from the original cultures in various morphological, cultural and physiological properties.

Among the variations reported for actinomycetes, the lytic activities of certain strains deserve attention. Dmitrieff and Souteeff[1] studied a culture apparently belonging to the *Streptomyces* group and designated as *Actinomyces bovis* Bostroem. This culture underwent lysis in liquid broth and on solid media. On agar media, lysis was associated only with a certain type of colony; this phenomenon was always accompanied by the formation of daughter colonies. Among these, two types were recognized: one did not differ from the mother colony and preserved the capacity for lysis; the second did not lyse and differed from the first in morphology. The lysing forms possessed stronger proteolytic properties and apparently did not produce any aerial mycelium; the non-lysing colonies were less proteolytic, formed a chalky aerial mycelium, and changed the reaction of litmus milk to alkaline. Lysis took place in broth cultures in 2 to 3 weeks. It was associated with the living culture and was of the nature of a non-enzymatic and non-transmissible lytic factor.

Other investigators[3] as well found that the lytic factor of actinomycetes is not comparable to bacteriophage or to lysozyme, since it dissolved both dead and living cells, and was strictly species specific; it was also thermostable. The fact that lysis may be limited to certain actinomycetes only when grown upon special media was brought out by Katznelson[2] for a thermophilic organism. This organism changed the reaction of the medium to acid, the addition of $CaCO_3$ to the medium inhibiting lysis. This lytic agent was also non-transmissible.

Experimental.—In these experiments, a strain of *A. griseus* found[6] to be capable of producing the antibiotic substance streptomycin was used. In the hope that strains having greater antibiotic potency than the original culture might be obtained, an effort was made to isolate distinct colonies. Spore suspensions of the organism were plated out on different media, with the dilutions sufficiently high to allow the development of only a few isolated colonies per plate. Incubation took place at 28°C. for 3 to 6 weeks.

The individual colonies developing on the plates showed considerable variation. Some were wrinkled, others were smooth; some produced pure white aerial mycelium, but the majority were characterized by the typical grayish-green color of the mycelium. A few of the colonies were almost completely free of aerial mycelium, and were somewhat moist and glisten-

ing in surface appearance. In some cases sectors were produced which differed from the rest of the colony. Upon transfer to fresh agar media, certain sectors gave rise to growth having a rough or a smooth surface of the typical sporulating aerial mycelium; other sectors, however, were entirely free of aerial mycelium, and produced cultures stable in this respect. The color variation of the aerial mycelium did not appear to be very significant. Some of the sectors from white colonies gave, on further cultivation, the typical gray-green color characteristic of the aerial mycelium of *A. griseus;* in other cases, the color change was reversed.

The various types of colonies and some of the different sectors were transferred to agar media. A number of cultures were thus obtained. These were inoculated into flasks containing nutrient glucose broth, and incubated at 28°C. Some flasks were shaken continuously in order to produce submerged growth, whereas others were kept stationary. The antibiotic activity of the filtrates was measured after different periods of incubation.

The results presented in table 1, as well as other data not reported here,

TABLE 1

PRODUCTION OF STREPTOMYCIN BY DIFFERENT COLONIES AND SECTOR ISOLATES OF *A. griseus*

CULTURE NO.	ORIGIN OF CULTURE	UNITS/ML. OF FILTRATE, AFTER DAYS — SHAKEN CULTURES 3	5	6	8	STATIONARY CULTURES 8	12	14
1	Typical sporulating colony	30	44	55	25	98	214	203
2	White sector of above colony	34	36	45	36	64	162	128
4	Non-sporulating sector of a typical colony	0	0	0	0	0	0	0
7	Large pure gray colony	31	30	83	33	45	136	135

indicate that in general the sporulating colonies, whether producing white, gray or gray-green mycelium, were active producers of streptomycin. The non-sporulating strains obtained from asporogenous sectors, however, were inactive. These non-sporulating strains produced no appreciable surface growth in stationary liquid cultures. The glucose consumption was much less, as compared with the active strains. The non-sporulating strains gave a lower pH value of the medium, namely, 5.0 to 6.5, whereas the filtrates of the sporulating types were always alkaline, ranging from pH 7.5 to 8.5. These and other physiological differences pointed to the fact that the non-sporulating forms were markedly different from the sporulating cultures. The two could have been considered as distinct species, had they been isolated separately from a natural substrate.

There was still a possibility that the inactive strains may not have lost the capacity of producing streptomycin, but that this was neutralized by the simultaneous elaboration of an inhibiting substance. The following

experiments tried to eliminate such possibility. Two samples of purified streptomycin of known activity were dissolved, one in water and the other in an equal amount of a glucose-free culture filtrate of an inactive strain. These two solutions of streptomycin were tested for activity after a short period of incubation and found to be exactly the same, whether water or the culture filtrate of an inactive strain was used as a solvent and diluent. In a second experiment, stationary cultures of inactive strains were reinoculated after different periods of incubation with spores of active strains. The rate of glucose consumption and the pH were found to rise rapidly, accompanied by the formation of streptomycin, as shown in table 2. These

TABLE 2

PRODUCTION OF STREPTOMYCIN BY MIXTURES OF INACTIVE AND ACTIVE STRAINS OF *A. griseus*

CULTURE NO.[a]	UNITS/ML. OF FILTRATE, IN DAYS INCUBATION 5	9	11	14	4[b]	6[b]	9[b]	FINAL pH
3	0	0	0	0	..	..	...	5.0
4	0	0	0	0	..	..	...	5.0
3 + 5[c]	83	88	128	86	..	..	...	...
4 + 5[c]	65	85	128	81	..	..	...	...
5	25	48	100	90	..	..	...	8.4
3, followed by 5	..	..	...	..	29	90	128	8.4
4, followed by 5	..	..	...	..	25	90	133	8.4

[a] Cultures 3 and 4 were inactive strains obtained from non-sporulating sectors; culture 5 was an active strain.

[b] Days after reinoculation.

[c] 3 + 5 or 4 + 5 simultaneous inoculation with cultures 3 and 5 or 4 and 5, respectively.

data also reveal that the simultaneous inoculation of a medium with an active and an inactive strain gave virtually the same antibiotic activity as inoculation with an active strain alone. In other words, the inactive strains had apparently no inhibitory effect upon streptomycin production by the active strains or upon the action of streptomycin itself.

The possibility that the initial acidity produced by the inactive strains might inactivate any streptomycin formed was also eliminated, as indicated by the results of another experiment reported in table 3. Addition of the buffering substances, $CaCO_3$ and K_2HPO_4, prevented an acid reaction, but did not result in the formation of streptomycin.

A study was made next of the cycle of growth of the inactive strains in shaken and in stationary cultures. The shaken cultures produced large balls of growth after 2 to 3 days' incubation; however, such cultures gradually developed into a fine, flocculent type of growth that is characteristic of the active strains grown in shaken culture from spore inoculation. The shaken cultures of the inactive strains finally underwent, after 6 to 9 days of incubation, a fairly rapid and complete lysis. It took about twice as long for the submerged growth of active strains to undergo lysis. In the stationary cultures of the inactive strains, no rapid lysis occurred;

TABLE 3

EFFECT OF BUFFER UPON THE PRODUCTION OF STREPTOMYCIN, SHAKEN CULTURES, 4 DAYS OF INCUBATION

STRAIN USED	GLUCOSE LEFT,[a] MG.	STREPTOMYCIN UNITS /ML.	pH
No. 19	0	34	7.8
No. 19 + variant 4	0	34	7.9
Variant 4	175	0	5.2
Variant 4 + $CaCO_3$[b]	0	0	7.4
Variant 4 + K_2HPO_4[b]	44	0	7.1

[a] 1 gm. in control.
[b] 5 gm. per liter.

these cultures consisted of submerged flocculent masses of growth on the bottom of the flask often accompanied by a surface ring around the glass wall. This type of growth persisted even after 30 days' incubation. Upon inoculation of stationary cultures of inactive strains with spores of an active strain, the submerged flocculent growth of the former underwent lysis as the characteristic pellicle of the latter developed.

When the vegetative growth taken from the subsurface portions of colonies of the active strains growing upon an agar plate was inoculated into stationary flasks, the submerged type of growth was obtained. It was similar in every respect to growth of an inactive strain; the pH of the medium was lowered, the glucose consumption was slow, and the filtrate was inactive. Upon the reinoculation of such cultures with spores of an active strain, the pH and rate of glucose consumption rose; the filtrate became active, and the submerged growth gradually lysed as a pellicle developed. An active strain was thus induced to behave like an inactive, non-sporulating strain when the vegetative growth was used as the inoculum. Somewhat analogous results were obtained in shaken cultures when the vegetative growth was used as the inoculum.

To determine whether the production of streptomycin is associated with the sporulating form of the organism, an attempt was made to reisolate sporulating strains, that is, strains producing aerial mycelium, from the non-sporulating variants. These reverted strains were obtained from an occasional old stationary culture of an inactive strain, which developed spontaneously one or more colonies producing sporulating sectors. Transfers made from such sectors yielded cultures which sporulated and produced streptomycin.

It has been shown elsewhere[7] that of various actinomycetes, the streptomycin-producing strain of *A. griseus* is most resistant to the effect of this substance. This method of approach was utilized for the purpose of establishing any genetic differences between the non-sporulating and the sporulating strains of *A. griseus*. For purposes of comparison, several independently isolated cultures of this organism were compared with the

various strains obtained from the streptomycin-producing culture (table 4). The two old cultures of *A. griseus*, namely, Nos. 3326*a* and 3378 not producing any streptomycin, were found to be highly sensitive to the antibiotic effect of this material. On the other hand, the streptomycin-producing culture of *A. griseus*, as typified by the active strains Nos. 4 and 19, were highly resistant to the action of this antibiotic. However, the non-sporulating, inactive variants Nos. 3, 4 and 6 were sensitive to streptomycin. The reverted sporulating form obtained from the non-sporulating variant produced streptomycin and was resistant to its action.

Of particular interest in the above experiment, is variant 6. This variant is intermediate with respect to the active sporulating strains and the inactive non-sporulating strains. In its growth on agar slants it slowly forms a limited amount of aerial mycelium. In liquid media, it produces at first a typical submerged growth characteristic of the non-sporulating strains, and subsequently forms some surface growth which tends to sporulate, although less rapidly and less abundantly than the active sporulating strains. It falls between the two in its capacity to produce a small amount of streptomycin, which results only after the surface growth begins. Finally, it is somewhat more resistant than the true non-sporulating variants 3 and 4 to the antibiotic action of streptomycin. Variant 6 thus combines the capacities of both the sporulating and the non-sporulating strains, being intermediate in its cultural and biochemical characteristics.

TABLE 4

EFFECT OF STREPTOMYCIN UPON THE GROWTH OF DIFFERENT STRAINS OF *A. griseus* AND UPON VARIANTS OBTAINED FROM THE SAME STRAIN

CULTURE OR STRAIN OF *A. griseus*	ORIGIN	PRODUCTION OF STREPTOMYCIN[a]	INHIBITION OF GROWTH[b]
Culture 3326*a*	Original isolation of 1916	0	<3.1
Culture 3378	Later isolate from soil	0	<3.1
Strain No. 4	Sporulating active form	38	>3, 125
Strain No. 19	Sporulating active form	128	>3, 125
Variant 3	Non-sporulating form	0	19.7
Variant 4	Non-sporulating form	0	15.6
Variant 6	Non-sporulating form	4	27.2
Reverted strain from variant 4	Sporulating active form	37	>3, 125

[a] Units of streptomycin in 12-day cultures.

[b] Units of streptomycin required to inhibit growth in 1 ml. of medium.

A summary of the properties of the active strain and the inactive variant of *A. griseus* is presented in table 5.

Discussion.—The results of experiments on the variation of a certain strain of *Actinomyces griseus* are reported. A certain strain of this organism, having the capacity of producing the antibiotic substance streptomycin, was separated into at least two types; these varied in their mor-

phology, namely, formation of an aerial mycelium, and in their physiology, such as production of antibiotic substance, formation of acid, rate of glucose consumption, autolysis, and production of a gum-like material. Between these two extremes intermediary strains were obtained from which either of the first two types could be readily isolated.

It is to be recalled, in this respect, that the major basis for the separation

TABLE 5

CULTURAL AND PHYSIOLOGICAL CHARACTERISTICS OF THE STREPTOMYCIN-PRODUCING STRAIN OF *A. griseus* AND ITS INACTIVE VARIANT

ACTIVE STRAIN	INACTIVE VARIANT
1. *Antibiotic activity.* Produces streptomycin in both shaken and stationary cultures.	1. No streptomycin formed either in shaken or stationary cultures.
2. *Growth.* Surface growth always heavily sporulated; grayish-green aerial mycelium.	2. No sporulating aerial mycelium; scant development of aerial hyphae with slight tendency to form spores in some old cultures.
3. *Reaction.* Medium always changes to alkaline; pH 7.5–8.5.	3. Reaction of medium at first acid, pH 5.0–6.5; later becoming alkaline.
4. *Glucose.* Glucose completely consumed in 6–8 days in stationary cultures and in 3–4 days in shaken cultures.	4. Glucose utilized more slowly.
5. *Lysis in shaken cultures.* Shaken cultures produce very fine flocculant growth, tending to lyse slowly after about 15 days.	5. Cultures produce at first balls of growth which change into the turbid, flocculant type; rapid and complete lysis in 7–10 days.
6. *Lysis in stationary cultures.* Surface pellicles stable; any submerged, flocculant growth tends to lyse as the surface pellicle develops.	6. Stationary cultures produce no surface growth but flocculant, submerged mycelial growth which lyses slowly, only after a month or longer.
7. *Viscosity.* Culture filtrate not showing any viscosity.	7. Culture filtrate becomes viscous during or after lysis.
8. *Reinoculation.* Inoculation of cultures with lysed inactive culture induces no lysis or reduction in activity.	8. Inoculation of cultures with spores of active strain produces growth and antibiotic activity if some glucose remains.
9. *Variation.* Sporulating strain gives rise to non-sporulating variants.	9. Asporogenous variants may reconvert to active, sporogenous forms.
10. *Sensitivity to streptomycin.* Very resistant to this antibiotic.	10. Very sensitive to this antibiotic.

of two of the four important genera of the *Actinomycetaceae, Streptomyces* and *Nocardia,* is the production of an aerial mycelium: the first genus produces such mycelium, whereas the second produces none or only a limited quantity.[9] The fact that a single strain representing a typical *Streptomyces,* namely *A. griseus,* can be made to yield a mutant or variant which has all the characteristics of a *Nocardia* suggests the logical question:

To what extent do the various species of *Nocardia* isolated from natural substrates represent degenerate species or accidental forms of the *Streptomyces* group?

Other significant facts brought out in these studies are the marked physiological properties that characterize different types of actinomycetes. This is contrary to the general assumption in regard to the paucity of certain stable physiological reactions for the characterization of these organisms. A freshly isolated strain of *A. griseus* having the capacity of producing streptomycin formed typical aerial mycelium, characteristic of the species. It changed the reaction of a glucose-containing medium to alkaline, produced characteristic types of surface and submerged growth, underwent only limited lysis, and was markedly resistant to the antibiotic action of streptomycin. On the other hand, the non-sporulating variant produced no aerial mycelium, formed no streptomycin, was sensitive to the antibiotic action of this substance, was characterized by typical growth, which in shaken culture underwent rapid lysis, and produced acid in the glucose-containing medium. Both strains possessed otherwise the various cultural properties which are characteristic of the *A. griseus* species as a whole, such as lack of pigmentation in organic media and also proteolytic and diastatic properties. The non-sporulating strain, when isolated as such, however, would hardly be recognizable as typical *A. griseus*.

Summary.—A strain of *Actinomyces griseus* characterized by the production of the antibiotic substance streptomycin was found capable of variation. Some of the variants produced no aerial mycelium and no streptomycin and were characterized by other properties which distinguished them from the original culture.

Sporulating and streptomycin-producing strains comparable in many respects to the original culture could be isolated under certain conditions of cultivation from the non-sporulating variants.

The original culture of *A. griseus* had all the properties of the newly created genus *Streptomyces;* however, the variant could be classified within the genus *Nocardia*.

The question is, therefore, raised: to what extent may many of the species of *Nocardia* described in the literature represent variants of *Streptomyces* species that have lost the property of producing aerial mycelium?

* Journal Series Paper, New Jersey Agricultural Experiment Station, Rutgers University, Department of Microbiology.

[1] Dmitrieff, S., and Souteeff, G., *Ann. Inst. Past.*, **56**, 470–476 (1936).

[2] Katznelson, H., *Soil Sci.*, **49**, 83–89 (1940).

[3] Krasilnikov, N. A., and Koreniako, A. I., *Microbiologia* (*U.S.S.R.*), **7**, 837 (1938).

[4] Kriss, A. E., "Variability of Actinomycetes," *Academy Sciences, U.S.S.R., Moscow* (1937).

[5] Lieske, R., "Morphologie und Biologie der Strahlenpilze," *Gebr. Borntraeger, Leipzig* (1921).
[6] Schatz, A., Bugie, E., and Waksman, S. A., *Proc. Soc. Exp. Biol. Med.*, **55**, 66–69 (1944).
[7] Schatz, A., and Waksman, S. A., *Ibid.*, **57**, 244–248 (1944).
[8] Waksman, S. A., *Soil Sci.*, **8**, 71–215 (1919).
[9] Waksman, S. A., and Henrici, A. T., *Jour. Bact.*, **46**, 337–341 (1943).

ON EPINEPHRINE AND RETINAL PHOTOMECHANICAL RESPONSES

BY S. R. DETWILER

DEPARTMENT OF ANATOMY, COLLEGE OF PHYSICIANS AND SURGEONS, COLUMBIA UNIVERSITY

COMMUNICATED MARCH 30, 1945

When the eyes of many vertebrates (particularly fishes, amphibians and birds) are exposed to light, the retinal pigment migrates, the cones contract and the rods elongate. In darkness the inverse changes occur, viz., contraction of the pigment, elongation of the cones and retraction of the rods. These so-called photomechanical responses (especially in the lower forms) have been regarded as important in adapting the retina to changes in illumination.[1–4] In man and other mammals where these responses apparently fail, adaptation to changing illumination must be sought primarily in the physiology of the retinal photopigments in combination with pupillary responses.

Photomechanical responses have been shown to be influenced by factors other than light and darkness.[2] An important factor, among others, is temperature. Both in fishes and in amphibia, low and high temperatures in the dark will bring about a rather characteristic light response.[2, 5, 6]

Numerous investigators[7–11] have shown that the injection of commercial epinephrine hydrochloride into dark-adapted frogs will produce an expansion of the pigment (light condition) in the dark. This response has been brought into question by Nover,[12] who claimed that, in reality, the migration of the pigment in the dark is not due to the hormone itself but to its acidity. In this respect he supported the acid stimulation theory of Dittler[13] and its amplification by v. Studnitz.[14] The latter maintains that light breaks down cone-photosensitive substance with the production of phosphoric acid, and that the acid initiates the pigment migration and positional changes in the photoreceptors. This theory has received ardent support by Wigger[15] and Nover,[12] and the whole matter of retinal photomechanical shifts is brought into direct dependence upon the H-ion con-

centration (acidity producing pigment expansion and cone contraction; alkalinity producing the inverse changes).

Nover[12] claims that when commercial adrenalin ("Suprarenin Hydrochlor.") is neutralized to pH 6.98 with NaOH and injected into frogs, it has no more effect than the injection of saline solution of the same pH. Consequently the hormone is regarded as ineffectual.

The results of the experiments herewith presented do not support Nover's findings. These experiments were performed upon frogs which had been

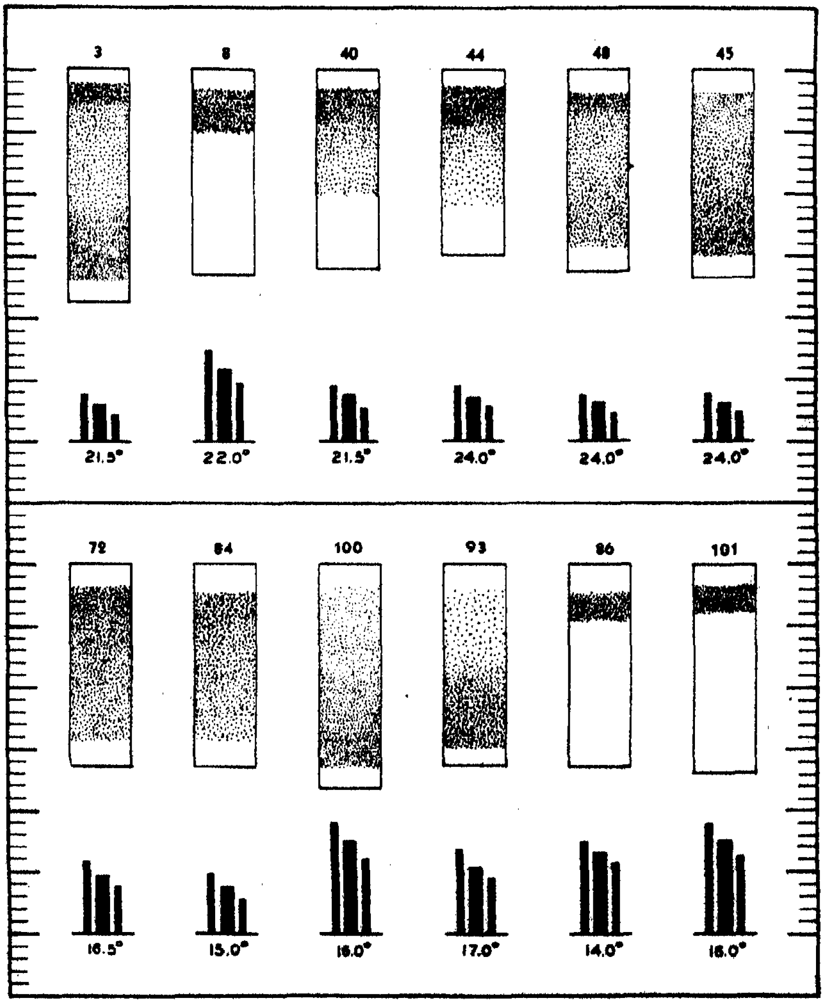


FIGURE 1

(See opposite page for explanation)

kept in darkness for approximately 2 hours—a period sufficient to produce a contracted state of the pigment and elongation of the cones (Fig. 1, case 8). After dark-adaptation the frogs were injected with epinephrine hydrochloride ("Adrenalin"; Parke, Davis and Co.) under conditions stated in table 1. The results upon the pigment and cones are given in figure 1. It becomes clear from a study of these results that epinephrine, whether markedly acid (pH 3.6–4.8) or brought up to neutral or slightly beyond with carbonate (pH 7–7.4), produces a marked migration of the pigment in the dark. Furthermore, injections of the pure base suspended in peanut oil produced the most striking pigment expansion. In most of

EXPLANATION OF FIGURE 1

Graphic representation of the responses of the frog retinal pigment and cones in various experiments, the protocols of which are given below. The data are partially listed in table 1. The vertical rectangles represent segments of the retina from the base of the epithelial cell (above) to the external limiting membrane (below). The pigment is represented by dots, which show the extent of migration or contraction, as well as density. The lengths of the cones are represented by the vertical solid black columns. The left narrow column for each eye represents the lengths of the longest cones (principal member of double cones); the right narrow column represents the lengths of the shortest cones (single cones). The broad middle column represents the mean length based upon 20 measurements for each retina (10 double and 10 single). Cone measurements are based on the distance in μ from the external limiting membrane to the outer segment (distal margin of the oil globule). Each space on the ordinates equals 4 μ.

Frog 3. Light-adapted 3 hours.

Frog. 8. Dark-adapted 2 hours.

Frog 40. Dark-adapted 2 hours. Injected with 1 cc. of Ringer's solution (pH 6.5). Re-dark-adapted 40 minutes. Eyes excised.

Frog 44. Dark-adapted 2 hours. Injected with 1 cc. of Ringer's solution (pH 6.5). Re-dark-adapted 20 minutes. Eyes excised.

Frog 48. Dark-adapted 2 hours, 30 minutes. Injected with 1 cc. of epinephrine hydrochloride (1:100,000). Re-dark-adapted 30 minutes. Eyes excised.

Frog 45. Dark-adapted 2 hours. Injected with 1 cc. of epinephrine hydrochloride (1:100,000). Re-dark-adapted 20 minutes. Eyes excised.

Frog 72. Dark-adapted 2 hours, 40 minutes. Injected with 1 cc. of epinephrine hydrochloride (1:50,000), neutralized to pH 7.4. Re-dark-adapted 30 minutes. Eyes excised.

Frog 84. Dark-adapted 2 hours, 10 minutes. Injected with 1 cc. of epinephrine hydrochloride (1:50,000), neutralized to pH 7.3. Re-dark-adapted 40 minutes. Eyes excised.

Frog 100. Dark-adapted 2 hours, 15 minutes. Injected with 0.5 cc. of epinephrine in peanut oil (1:500) 1 mg. Re-dark-adapted 2 hours, 30 minutes. Eyes excised.

Frog 93. Dark-adapted 2 hours, 10 minutes. Injected with 0.5 cc. of epinephrine in peanut oil (1:500) 1 mg. Re-dark-adapted 2 hours, 30 minutes. Eyes excised.

Frog 86. Dark-adapted 2 hours, 10 minutes. Injected with 0.5 cc. of peanut oil. Re-dark-adapted 3 hours. Eyes excised.

Frog 101. Dark-adapted 2 hours, 10 minutes. Injected with 0.5 cc. of peanut oil. Re-dark-adapted 2 hours, 30 minutes. Eyes excised.

TABLE I

Showing Effects of Injection of Epinephrine Hydrochloride into the Ventral Lymph Sac of Dark-adapted Frogs, and Re-dark-adapting for Periods Indicated

Substance injected	pH of solution	Case	Re-dark-adapted	Temp. in deg. C	Extent of pigment migration	Length of cones*: Longest	Shortest	Mean
1 cc. of 1:100,000 sol. epineph. hydrochlor.	...	45	20 min.	24.0	++++	15.6	9.6	12.4
	...	48	30 min.	24.0	+++	15.6	9.6	12.7
	...	50	20 min.	24.0	+++	20.4	14.4	17.4
1 cc. of 1:50,000 sol. epineph. hydrochlor.	4.8	76	35 min.	16.5	++++	22.8	15.6	18.8
	3.6	85	35 min.	15.0	+++	..	..	..
	7.4	72	30 min.	16.5	+++	22.8	15.6	19.4
	7.3	83	40 min.	15.0	+++	18.0	10.8	14.8
	7.3	84	40 min.	15.0	+++	19.2	10.8	14.9
1 cc. of 1:25,000 sol. epineph. hydrochlor.	7.0	104	40 min.	13.0	+++	..	..	..
	7.0	105	40 min.	13.0	+++	25.2	16.8	21.4
Epinephrine suspension in peanut oil								
0.5 cc. of 1:500 (1 mg.)	Neutral	93	2 hrs. 30 min.	17.0	++++	26.4	18.0	21.7
		94	2 hrs. 30 min.	17.0	++++	30.0	19.8	23.9
		100	2 hrs. 30 min.	16.0	++++	36.0	24.0	30.2
0.25 cc. of 1:500 (0.5 mg.)		102	2 hrs. 30 min.	16.0	++++	34.8	21.6	29.1
		103	2 hrs. 30 min.	16.0	++++	39.6	25.2	34.3
0.5 cc. of peanut oil	Neutral	80	40 min.	14.5	+	22.8	14.4	18.5
		86	3 hrs.	14.0	–	30.0	22.8	26.4
		101	2 hrs. 30 min.	16.0	–	36.0	25.2	30.7

++++ = Extensive migration with distal accumulation of pigment (e.g., case 45, Fig. 1).
+++ = Extensive migration with no distal accumulation (e.g., case 48, Fig. 1).
++ = Pigment migrated between one-third and two-thirds down (e.g., case 40, Fig. 1).
+ = Very slight migration of pigment.
– = No migration. Typical dark condition (e.g., case 86, Fig. 1).

* Longest cones are principal members of double cones. Shortest cones are single cones. Mean length is based on average of 20 measurements for each case.

these latter cases the pigment became heavily massed near the external limiting membrane (e.g., case 93, Fig. 1). The cones, however, in many of these retinas were as long as those in the retinas of control frogs, i.e., those injected with peanut oil alone where, after $2^1/_2$ to 3 hours in darkness, the pigment was found to be maximally contracted and the cones elongated (Fig. 1, cf. cases 93 and 100 with cases 86 and 101). These results indicate clearly that pigment migration and cone contraction do not always go together.

It has been shown earlier[16] that any excitation of the frog in the dark will produce a partial light response of the retina in darkness. It has been suggested that this response may be due to the increased secretion of the adrenal medulla resulting from excitation. In the present experiments this tentative theory would seem to be supported inasmuch as injections of epinephrine produce marked pigmentary responses in the dark. That the adrenal gland exerts a definite influence on retinal pigmentary responses has been shown by Hasama.[17] He found that after cauterization of the frog adrenal gland, the pigment failed to expand upon exposure to light. Furthermore, faradic stimulation of the adrenals in dark-adapted frogs was shown to produce the light condition of the pigment.

Summary.—When commercial epinephrine hydrochloride ("Adrenalin"; Parke, Davis and Co.) is injected into dark-adapted frogs, the retinal pigment undergoes marked expansion in the dark. This ensues whether the solution is acid (pH 3.6–4.8) or is neutralized with carbonate to pH 7. Injections of the pure epinephrine base suspended in peanut oil produce maximal migration in $2^1/_2$–3 hours. When peanut oil alone is injected, the pigment, after equivalent exposures to darkness, is in a highly contracted state. The results indicate that the pigment expansion is due to the action of the hormone itself, and not the result of acid stimulation.[18]

[1] Arey, L. B., *Jour. Comp. Neur.*, **25**, 535–554 (1915).
[2] Arey, L. B., *Ibid.*, **26**, 121–201 (1916).
[3] Walls, G. L., "The Vertebrate Eye," Cranbrook Institute of Science, Bull. 19, Bloomfield Hills, Michigan, 1942.
[4] Detwiler, S. R., *Vertebrate Photoreceptors*. The Macmillan Co., New York, 1943.
[5] Herzog, H., *Arch. Anat. Physiol. Phys. Abt.*, pp. 413–464 (1905).
[6] Detwiler, S. R., and Lewis, R., *Jour. Comp. Neur.*, **41**, 153–169 (1926).
[7] Arey, L. B., *Anat. Rec.*, **16**, 138 (1919).
[8] Dubois-Poulsen, A., *Comp. Rend. Soc. Biol.*, **125**, 248–249 (1937).
[9] Drouet, P. L., and Florentin, P., *Rev. Médicale de Nancy*, **65**[2], 678–689 (1937).
[10] Shima, I., *Zeit. ges. exp. Med.*, **102**, 535–541 (1938).
[11] Sverdlick, J., *Rev. Soc. Argentina Biol.*, **28**, 207–214 (1942).
[12] Nover, L., *Pflüger's Arch.*, **242**, 665–684 (1937).
[13] Dittler, R., *Ibid.*, **117**, 295–328 (1907).
[14] v. Studnitz, G., "Physiologie des Sehens," Bd. 3 of *Probleme der Biologie*, Akad. Verlagsgesellschaft, Leipzig, 1940.

[15] Wigger, H., *Pflüger's Arch.*, **239**, 215–246 (1937).

[16] Detwiler, S. R., *Jour. Comp. Neur.*, **81**, 137–145 (1944).

[17] Hasama, B., *Pflüger's Arch.*, **238**, 758–766 (1937).

[18] A detailed account of the effect of injections of acids and alkalis upon retinal pigmentary responses in the frog will be published elsewhere.

THE VARIATION OF INFECTIVITY

BY EDWIN B. WILSON AND JANE WORCESTER

HARVARD SCHOOL OF PUBLIC HEALTH

Communicated April 10, 1945

There are two main approaches to the study of the epidemic curve of new cases. First is what may be called the dynamical approach in which *a priori* assumptions are made as to the laws regulating the relations between the changing case rates and the number of susceptibles; Ronald Ross, Soper, McKendrick and others have followed this line of attack. Second is what may be called the statistical approach in which some form is assumed for the curve of new cases, generally based *a posteriori* on the observation that such a curve does describe fairly well the course of new cases; Farr, Brownlee and others have pursued this method.

We desire to examine the two methods together to see what conclusions result. The three curves which have been most used are: (1) the probability curve

$$C = \frac{T}{\sqrt{2\pi}\,\sigma} e^{-\frac{t^2}{2\sigma^2}}, \tag{1}$$

where T is total cases during the whole epidemic, t is the time in any unit, σ the standard deviation of the distribution of the case rates in time, and C is the case rate of new cases at the time t; (2) Pearson's Type IV, much used by Brownlee in its general form, which we shall use only in its symmetrical form[1] (Type VII), namely,

$$C = \frac{T\Gamma(n)}{a\sqrt{\pi}\,\Gamma(n - {}^1/_2)}\left(1 + \frac{t^2}{a^2}\right)^{-n}, \tag{2}$$

where $a = (2n - 3)^{1/2}\sigma$ and Γ is the symbol of the gamma-function; (3) the derivative of the growth curve, namely

$$C = {}^1/_2 T\beta \operatorname{sech}^2 \beta t, \tag{3}$$

where $\beta = \pi/(2\sqrt{3}\,\sigma) = 0.9069/\sigma$. Each of these symmetrical curves has been referred to its mean.

The dynamical law which we shall use is the law of mass action as we have recently generalized it,[2] viz.,

$$C(t) = \left(\frac{S}{m}\right)^p C(t - \tau), \tag{4}$$

for a disease of definite incubation time τ. The number m, which is the number of susceptibles just sufficient so that the case rate is the same at t as at $t - \tau$, can be taken as *one* measure of the reciprocal of the infectivity. The infectivity thus defined is a strictly epidemiological concept and has no intimate connection with bacterial virulence. Given any curve of new cases, we have

$$m = S[C(t - \tau)/C(t)]^{1/p}. \tag{5}$$

The value of S is not known but if S_0 be the number of susceptibles at the peak of the epidemic curve

$$S = S_0 - \int_0^t C\,dt \tag{6}$$

and the values of S at the beginning and the end of the epidemic are

$$S_B = S_0 + T/2, \qquad S_E = S_0 - T/2 \tag{7}$$

and the value of m at the peak is very nearly

$$m_0 = S_0 - {}^1/_2 \text{ peak cases} \tag{8}$$

where peak cases are defined as the number of cases that would occur during the period τ if the peak case rate were maintained throughout that interval of time.

Brownlee defines infectivity[3] directly from the epidemic curve without reference to the number of susceptibles as

$$\frac{C + \Delta C}{C} = \frac{C(t + \tau)}{C(t)} = f(t). \tag{9}$$

From this it appears that, if $p = 1$ in the dynamical law, his definition makes

$$f(t) = \frac{S(t - \tau)}{m(t - \tau)}, \qquad \text{or } f(t + \tau) = \left(\frac{S}{m}\right)^p \tag{10}$$

in the general case. Actually, like many workers, he assumes that τ is short enough so that differentials may replace increments and

$$f(t) = e^{\tau(d \log C/dt)}. \tag{11}$$

If the three curves (1)–(3) be taken as the epidemic curve, the infectivity as defined by Brownlee in (11) gives, respectively,

$$f(t) = e^{-\tau t/\sigma^2} \quad \text{or } e^{-\frac{2n\tau t}{a^2 + t^2}} \quad \text{or } e^{-2\beta\tau \tanh \beta t}.$$

For t large and negative the results are ∞, 1, $e^{2\beta\tau}$; whereas for t large and positive they are 0, 1, $e^{-2\beta\tau}$; and at the peak of the epidemic ($t = 0$) they are[4] 1, 1, 1.

If m be taken as a measure of the reciprocal of the infectivity and equation (5) be used, the results are somewhat more complicated but not too dissimilar. One may begin with (3), whence

$$m = \frac{\operatorname{sech}^{2/p} \beta\tau(S_0 - {}^1/_2 T \tanh \beta t)}{(1 - \tanh \beta\tau \tanh \beta t)^{2/p}}.$$

If $S_0 : T/2 = \coth \beta\tau$ the numerator and denominator become proportional and the value of m increases throughout the epidemic, i.e., the infectivity $1/m$ falls off, provided $p < 2$; but the opposite holds provided $p > 2$; and for $p = 2$ the value of m is constant. If, however, $S_0 : T/2 \neq \coth \beta\tau$ the relations become more complicated; in the special case $p = 2$ the value of m will increase throughout the epidemic if $S_0 : T/2 > \coth \beta\tau$ but will decrease if $S_0 : T/2 < \coth \beta\tau$.

If attention be turned to the normal curve (1),

$$m = \left(S_0 - \frac{T}{\sqrt{2\pi}\,\sigma} \int_0^t e^{-\frac{t_2}{2\sigma^2}} dt\right) e^{\tau(2t - \tau)/2p\sigma^2}.$$

When $t = -\infty$, $m = 0$; when $t = 0$, $m = S_0 e^{-t^2/2p\sigma^2}$; when $t = +\infty$, $m = \infty$ provided S_∞ does not become zero and it generally does not. Although m is greater when $t = 0$ than when $t = -\infty$ and less when $t = 0$ than when $t = \infty$ no matter what the value of t, the rate of change of m when $t = 0$ will be negative if

$$p > \sqrt{2\pi}\, \frac{\tau}{\sigma} \frac{S_0}{T} = 2.507\, \frac{\tau}{\sigma} \frac{S_0}{T}.$$

As S_0 need not be larger than $T/2$ and as τ/σ is generally less than $^1/_2$ when the epidemic runs for a considerable number of incubation periods, it is clear that p need not be very large to satisfy the relationship; and when the relationship is satisfied m starts at 0 when $t = -\infty$, increases to a maximum prior to the peak of the epidemic, decreases to a minimum after the peak, and then increases indefinitely.

The sort of results that arise on combining the generalized law of mass action with specific assumptions as to the type of epidemic curve which may be assumed "as found by trial apart from theory," including that particular type which has been shown to arise by theory as a first approximation (when there are no recruits), may be exhibited better by a specific illustrative example than by detailed development of formulas. Let us take the 1925–1926 epidemic from Hedrich's Baltimore series. Table 1 gives the estimated cases by months, the cases that would be calculated from curves of types (1), (2), (3) fitted to the data, and the values of the infectivity taken as $70604/m$ computed from these curves (not from the data) for the hypotheses $p = 1, 2, 4$. The value assumed for S_0 at the mode was 72982 and the equilibrium value was figured as 70604. It will

TABLE 1

THE VARIATION OF THE INFECTIVITY UNDER A VARIETY OF HYPOTHESES

TIME	CASES	(1)[3]	(2)[4]	(3)[5]	VALUE[6] OF $\frac{70604}{m}$ FROM (1)				VALUE OF $\frac{70604}{m}$ FROM (2)				VALUE OF $\frac{70604}{m}$ FROM (3)			
					$p = 1$	$p = 2$	$p = 4$	B[7]	$p = 1$	$p = 2$	$p = 4$	B	$p = 1$	$p = 2$	$p = 4$	B
$-\infty$	...[2]	0	0	0	∞	∞	∞	∞	0.80	0.80	0.80	1.00	1.49	1.09	0.93	1.87
Aug.	50	3	24	27	3.24	1.61	1.13	4.48	0.90	0.85	0.82	1.65	1.49	1.09	0.93	1.87
Sept.	72	44	104	103	2.60	1.44	1.07	3.53	0.91	0.85	0.82	1.73	1.48	1.09	0.93	1.87
Oct.	222	308	326	354	2.35	1.37	1.04	2.80	0.92	0.86	0.83	1.81	1.48	1.09	0.93	1.86
Nov.	1379	1412	1127	1177	1.86	1.22	0.99	2.20	0.93	0.86	0.83	1.86	1.47	1.08	0.93	1.83
Dec.	2602	4056	3490	3505	1.49	1.10	0.94	1.73	0.94	0.87	0.84	1.79	1.44	1.08	0.94	1.75
Jan.[1]	8370	7390	7748	7682	1.23	1.02	0.93	1.36	0.95	0.89	0.87	1.53	1.35	1.07	0.95	1.52
Feb.[1]	10931	8492	9540	9496	1.07	1.00	0.96	1.08	0.98	0.95	0.94	1.12	1.14	1.03	0.98	1.12
Mar.[1]	5060	6188	6022	5981	0.95	1.00	1.02	0.85	1.02	1.04	1.05	0.78	0.90	0.98	1.02	0.77
Apr.	1881	2843	2304	2344	0.82	0.98	1.06	0.67	1.04	1.10	1.14	0.60	0.75	0.93	1.04	0.61
May	708	825	705	742	0.68	0.91	1.05	0.53	1.05	1.13	1.17	0.55	0.69	0.91	1.05	0.56
June	209	152	206	219	0.54	0.82	1.00	0.41	1.05	1.14	1.18	0.54	0.67	0.91	1.06	0.54
July	168	19	66	64	0.43	0.73	0.95	0.33	1.06	1.15	1.19	0.56	0.66	0.91	1.06	0.54
Aug.	81	0	20	21	0.35	0.66	0.90	0.26	1.08	1.16	1.19	0.59	0.66	0.91	1.06	0.54
$+\infty$	...[2]	0	0	0	0	0	0	0	1.23	1.23	1.23	1.00	0.66	0.91	1.06	0.53

[1] The months of Jan., Feb., Mar. were adjusted to 30 days; the unit of time was taken as a half-month.

[2] The actual epidemic cannot well be followed past the epidemic year in either direction.

[3] The fitted curve (1) was taken to be $C = 4370\, e^{-t^2/17.78}$ with mode about Feb. 9.

[4] The fitted curve (2) was taken to be $C = 4990\,(1 + t^2/41.95)^{-4}$ with mode at the same time.

[5] The fitted curve (3) was taken to be $C = 4968\, \mathrm{sech}^2\,(0.3131t)$ with mode at the same time.

[6] The values are those for the first of the month; the value of 70604 is the "equilibrium" number m_0.

[7] The infectivities under B are those computed from (11) rather than from (9).

be observed that for each type of curve the fit to the data is rather poor and that, therefore, values of the infectivity calculated directly from the original data would undoubtedly behave differently from those based on any of the fitted curves. However, this table is offered for the purpose of showing how the infectivity would theoretically vary if these types of curve were supposed to hold precisely for an epidemic.

The following comments may be made on Table 1. Brownlee's infectivity decreases for (1) and (3) throughout the course of the epidemic curve, but for (2) it shows at the very ends (August) the reversal which has been noted in the theoretical discussion.[4] For the normal curve (1) the infectivity defined as proportional to the reciprocal of m and standardized to be unity when S has its equilibrium value m_0 and $C(t) = C(t - \tau)$, is highly variable if $p = 1$ or $p = 2$ (but only moderately variable if $p = 4$) within the natural limits of the epidemic; it decreases throughout the epidemic if $p = 1$ or $p = 2$ but shows the fluctuation near the mode if $p = 4$ as described in the text. For the curve (2) the infectivity increases throughout the epidemic being not very variable for $p = 1$ but increasingly so for $p = 2$ and $p = 4$. For the curve (3) there is a decreasing infectivity with considerable variation if $p = 1$ and with little variation if $p = 2$, whereas if $p = 4$ there is an increasing infectivity with slight variation; clearly for some value of p intermediate between 2 and 4 there would be practical (though not quite absolute) constancy in the infectivity. It is certain that whether the infectivity be defined as by Brownlee or whether it be taken from the generalized law of mass action, its quantitative (and even its qualitative) behavior is very different according to the type of epidemic curve selected.

[1] *Encyclopedia Britannica*, 14th ed., Vol. 8, p. 650. In this article on Epidemiology, Brownlee and Greenwood state: "The equation of the curve which describes the majority of epidemics, as found by trial apart from theory, is $y = a(1 + t^2/b^2)^{-n}$—(Hypothetically) the organism may be assumed to possess at the beginning of the disease a high degree of infectivity which decreases as the epidemic goes on." We may observe that Brownlee fits Type IV from four moments of the epidemic curve. There is no difficulty about this if the epidemic is that of a disease which starts and stops abruptly but there are real difficulties in finding reliable results for third and fourth moments in many cases where the beginning and end of the epidemic are somewhat indefinite. Further, for Type VII, $\beta_2 > 3$, the relation between n and β_2 being

$$n = (5\beta_2 - 9)/(2\beta_2 - 6) \qquad \text{or } \beta_2 = (6n - 9)/(2n - 5)$$

so that for $n = \infty$ we have $\beta_2 = 3$ (the normal curve) and n must be greater than $^5/_2$ for which β_2 would become infinite. For the derivative of the growth curve $\beta_2 = 4.2$ and Type VII fitted thereto by moments would need $n = 5$.

[2] These PROCEEDINGS, 31, 109–116 (1945).

[3] John Brownlee, "On the Curve of the Epidemic," *British Med. Jour.*, 1, 799–800 (1915).

[4] If (1) or (3) be used it is clear that $f(t)$ always decreases; but for (2)

$$f'(t) = 2n\tau e^{-\frac{2n\tau t}{a^2+t^2}}\left[\frac{t^2 - a^2}{(a^2 + t^2)^2}\right]$$

so that $f(t)$ increases from 1 when $t = -\infty$ to a maximum of $e^{n\tau/a}$ when $t = -a$, decreases to a minimum of $e^{-n\tau/a}$ when $t = +a$, and then increases to 1 when $t = +\infty$. The value of C as a fraction of its modal value C_0 at the time $t = a$ is 2^{-n} so that the infectivity according to Brownlee's definition is increasing up to the time when the case rate is $C_0/2^n$, then decreases to the time when again the case rate is $C_0/2^n$, and finally increases thereafter. If n is large the tails of the curve beyond the place where $C = C_0/2^n$ may be disregarded so that practically the statement of Brownlee and Greenwood that the infectivity decreases from beginning to end of the epidemic is true, but if n is large the statement cannot be quite true of Brownlee's definition and this epidemic curve.

GROUPS HAVING A SMALL NUMBER OF SETS OF CONJUGATE SUBGROUPS

By G. A. MILLER

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF ILLINOIS

Communicated April 5, 1945

The identity and the group itself are sets of conjugate subgroups of every group G. If G is the cyclic group of order p^m, p being an arbitrary prime number, then G contains exactly $m + 1$ sets of conjugate subgroups and when p is odd this cyclic group can be extended by p^m subgroups of order q, where q is an arbitrary prime divisor of $p - 1$, so as to obtain a group of order $p^m q$ which contains exactly $m + 3$ sets of conjugate subgroups. In particular, when $m = 1$ the group of order p contains exactly two sets of conjugate subgroups and this is the only group which contains exactly two sets of conjugate subgroups. The two groups of order pq contain separately exactly four sets of conjugate subgroups. One of these is cyclic and the other is non-cyclic and contains p subgroups of order q.

The cyclic group of order p^2 clearly contains exactly three sets of conjugate subgroups and this is the only group which contains exactly three sets of conjugate subgroups. This results from the facts that if a group contains exactly three sets of conjugate subgroups its order cannot be divisible by two distinct prime numbers and the central quotient group of every non-abelian prime power group is non-cyclic. If a group contains exactly four sets of conjugate subgroups its order cannot be divisible by more than two distinct prime numbers and when it is devisible by two distinct prime numbers the order is the product of these numbers.

From what precedes it results that if a group does not contain more than four sets of conjugate subgroups it is the identity if it contains one and

only one such set, it is a group of prime order if it contains two and only two such sets, it is the cyclic group whose order is the square of a prime number if it contains three and only three such sets, and it is either one of the two possible groups whose orders are the product of two distinct prime numbers or the cyclic group whose order is the cube of a prime number when it contains four and only four such sets. In all of these cases, except the first, the number of the possible groups is infinite and all of these groups are abelian except the groups of order pq which certain p subgroups of order q.

Before considering all the groups which contain five and only five sets of conjugate subgroups it may be desirable to consider the prime power groups with respect to the number of sets of conjugate subgroups. The non-cyclic group of order p^2 clearly contains $p + 3$ sets of conjugate subgroups and this is the smallest number of sets of conjugate subgroups in any non-cyclic prime power group. This is obvious when the group is abelian and when it is non-abelian it results from the facts that the central quotient group of a non-abelian group is always non-cyclic and that the sets of conjugate subgroups in a group is always larger than the number of sets of conjugate subgroups in any of its quotient group with respect to a non-identity invariant subgroup. By forming successive central quotient groups we arrive at an abelian group.

From what precedes it results that the only prime power non-cyclic group which contains as few as five sets of conjugate subgroups is the four group. The prime power cyclic group of order p^4 also contains exactly five sets of conjugate subgroups according to the theorem noted above. This is also true of the infinite system of groups of order p^2q which are obtained by extending the cyclic group of order p^2, p being any odd prime number, by p^2 cyclic groups of order q, where q is an arbitrary prime divisor of $p - 1$. The smallest groups in the latter category are the dihedral group of order 18 and the dihedral group of order 50. The smallest group of this category which is not dihedral is of order 147 and is obtained by extending the cyclic group of order 49 by 49 subgroups of order 3. There is no other group whose order is divisible by as few as two distinct prime numbers in which there are exactly five sets of conjugate subgroups with the exception of the tetrahedral group.

This results almost directly from the fact that all such groups are known to be solvable and hence one arrives at the identity by forming successive commutator subgroups. It is also true that when the order of G is divisible by three distinct prime numbers then G cannot contain exactly five sets of conjugate subgroups. In fact, the order of G could then not be divisible by the square of one of any of these prime numbers and hence G would again be solvable. In this case it would contain an invariant subgroup whose order would be equal to the product of the two largest of these three

prime numbers. This invariant subgroup could not be non-cyclic and hence all of its operators besides the identity would be non-commutative with one of the remaining operators of G. It would therefore result that G would contain more than five sets of conjugate subgroups, which is contrary to the hypothesis. We have therefore considered all the possible groups which separately involve as few as five sets of conjugate subgroups.

If G has six sets of conjugate subgroups and its order is a power of a prime number this number could not exceed 3 when G is non-cyclic because $p + 3$ would then exceed 6. If it is cyclic it is of order p^5, where p is an arbitrary prime number in accord with the general theorem noted above. Since every non-cyclic group of order p^m_1 contains at least $p + 1$ subgroups of index p and these are invariant under the group it results that every non-cyclic group of order p^m contains at least $p + 1$ sets of conjugate subgroups under the group which are separately of index p. Moreover, every group of order p^m contains at least one invariant subgroup of order p. Hence it results that when a group of order 2^m contains exactly six sets of conjugate subgroups it can contain only one subgroup of order 2 and is, therefore, either cyclic or dicyclic. That is, the quaternion group is the only group of order 2^m which contains exactly six sets of conjugate subgroups and is non-cyclic.

If a non-cyclic group of order 3^m contains exactly six sets of conjugate subgroups then $m = 2$ and G is the non-cyclic group of order 9 in accord with the general theorems noted above. That is, *if a prime power group contains exactly six sets of conjugate subgroups it may be the group of order p^5, p being an arbitrary prime number, the quaternion group, or the non-cyclic group of order 9, but it can be no other group.* It remains to consider the possible cases when G contains exactly six sets of conjugate subgroups but its order is not a power of a prime number.

The number of the distinct prime numbers which divide the order of G could clearly not exceed four since the number of the sets of conjugate subgroups of G is supposed to be six. If this former number would be four the order of G would not be divisible by the square of a prime number and hence the main properties of G would be known. In particular, G would have to contain at least two operators of prime orders which would be commutative. As this is impossible the order of G could not be divisible by as many as four distinct prime numbers when it contains exactly six sets of conjugate subgroups.

Suppose that the order of G is divisible by three distinct prime numbers but that the square of one of these prime numbers also divides the order of G. Although the simple group of order 60 satisfies this condition it is easy to prove that if we add the condition that the operators of G are contained in six sets of conjugates then G must be solvable and hence it must contain an invariant subgroup of prime index. The order of this

invariant subgroup could not be divisible by three distinct prime numbers since it would then contain a characteristic subgroup whose order would be the product of the two largest of these prime numbers and hence G would involve subgroups of more than six different orders, which is contrary to the hypothesis. The said invariant subgroup of G could also not be divisible by the square of a prime number for similar reasons.

If the order of G is divisible by three distinct prime numbers but not by the square of one of these numbers it contains an invariant subgroup whose order is the product of the two largest of these three prime numbers. If this invariant subgroup is cyclic it can clearly be extended by an operator which transforms it into itself but is not commutative with any operator of this invariant subgroup besides the identity. In particular, if the order of such a group is even there is always an operator of order 2 in the group which has this property. The group thus obtained clearly contains exactly six sets of conjugate subgroups and the dihedral group of order 30 is the group of smallest order which satisfies these conditions. When the order of G is pq^3 and G involves an invariant subgroup of order q^3 and of type 1^3 which involves no invariant operator besides the identity then G clearly involves exactly six sets of conjugate subgroups. The group of order 56 which contains eight subgroups of order 7 is a well-known illustration of the groups of this category.

The present article is closely related to the one published in these PROCEEDINGS, **30**, 359–362 (1944) under the heading "Groups containing a small number of sets of conjugate operators." It may be noted that in some cases the given number of conjugate subgroups imposes a greater restriction on the possible groups than the given number of conjugate operators since the number of the possible subgroups may be larger than the number of the possible operators in a given group.

THE VELOCITY OF PROPAGATION OF BRITTLE CRACKS IN STEEL

BY M. GREENFIELD AND G. HUDSON

DAVID TAYLOR MODEL BASIN, USN, WASHINGTON, D. C.

Communicated April 14, 1945

Recently there have been instances in which steel plate used in a variety of structures and subjected to certain severe conditions has ruptured in a brittle fashion. Characteristics of these brittle failures are (1) a crack is propagated with extreme rapidity through the steel plate, (2) the surface

of rupture created by this crack is orthogonal to the plane of the plate, (3) this rupture surface shows a characteristic "herringbone" pattern and (4) there exist small permanent strains, often of less than 2 per cent, in the neighborhood of the fractured surface.

Many engineering structures have failed in this catastrophic manner. One example which attracted a great deal of attention was the explosion at Schenectady, N. Y., of a spherical storage tank containing hydrogen.[1]

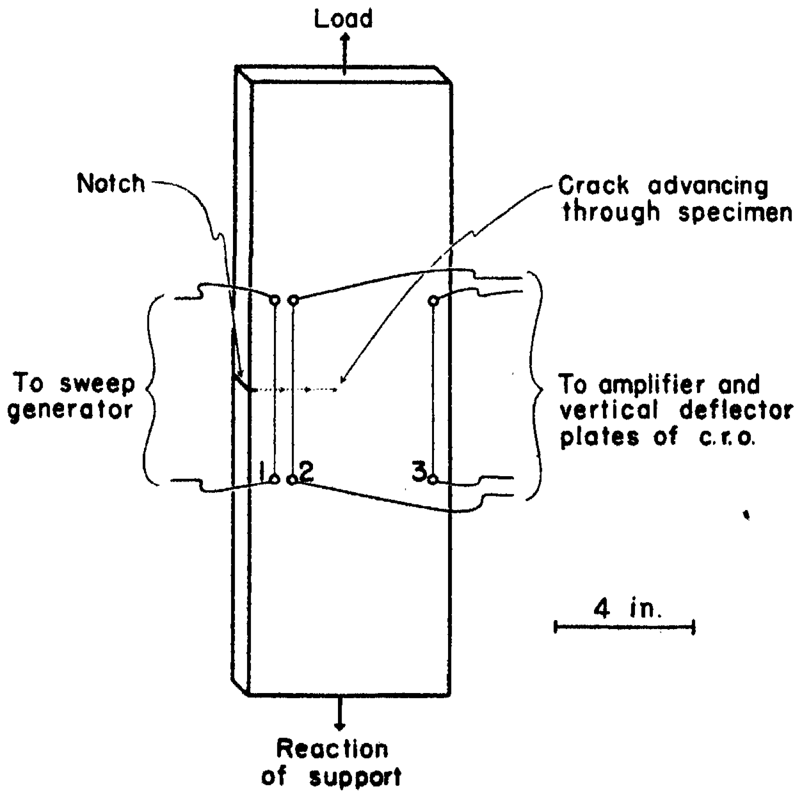


FIGURE 1

Schematic diagram showing the positions of the break-wires *1*, *2* and *3* on the tensile specimen.

Still another example is the breaking up of certain types of welded merchant vessels in a seaway. It is thought that a study of the speed of propagation of brittle cracks in steel will help in understanding the mechanism of these phenomena.

In this note we announce the results of some measurements which have been made of the velocity of propagation at room temperature of brittle

cracks in notched tensile specimens of medium steel. Several fine wires were cemented a few inches apart on the surface of a specimen, and placed normally to the expected path of the crack (Fig. 1). When the specimen is pulled to failure a crack is propagated from the notch across the specimen and breaks the wires in turn. The first wire which breaks triggers a sweep generator which moves a beam of electrons horizontally across the screen of a cathode ray oscillograph. As each succeeding wire breaks, the electron beam jumps vertically resulting in a record like that shown in figure 2.

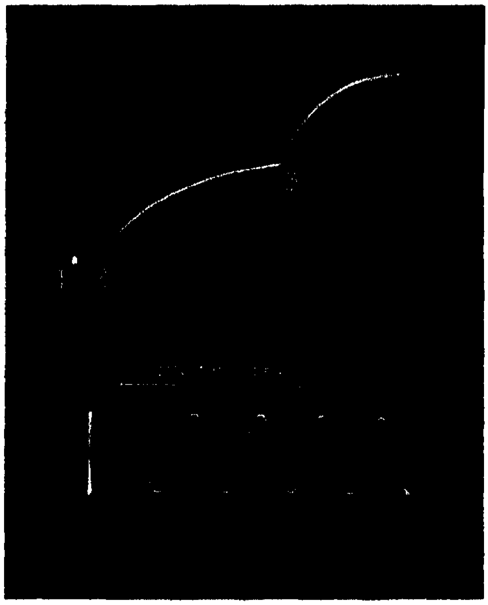


FIGURE 2

A typical record obtained from the breaking of a tensile specimen like that shown in figure 1. The numbers *1*, *2* and *3* designate the positions of the electron beam at the instant of breaking of wires *1*, *2* and *3*, respectively.

The time interval between jumps is measured from the record. This datum combined with the measured distance between the corresponding break-wires enables one to calculate the velocity of the crack. In this way there has been obtained an average velocity of 40×10^3 inches per second with a mean absolute deviation of 6 per cent.

It is expected that a paper giving more details of this work will be published in the near future.

[1] Brown, A. L., and Smith, J. B., *Mech. Eng.*, **66**, 392–397 (1944).

PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 31 June 15, 1945 Number 6

ON THE EVOLUTION OF BIOCHEMICAL SYNTHESES

BY N. H. HOROWITZ

SCHOOL OF BIOLOGICAL SCIENCES, STANFORD UNIVERSITY, CALIF.

Communicated April 23, 1945

Although it has been recognized for a long time that the biochemistry of the organism is conditioned by its genetic constitution, a more precise definition of this dependence has not been possible until recently. A considerable amount of evidence now exists for the view that there is a one-to-one correspondence between genes and biochemical reactions. This concept, foreshadowed in the work of Garrod[1] on human alcaptonuria, accounts in a satisfactory way for the inheritance of pigment formation in guinea pigs,[2] insects[3] and flowers,[4] and the synthesis of essential growth factors in *Neurospora*.[5] It appears from these studies that each synthesis is controlled by a set of non-allelic genes, each gene governing a different step in the synthesis. As to the nature of this control, it is probable that the primary action of the gene is concerned with enzyme production. That genes can direct the specificities of proteins has been shown in the case of many antigens,[6] while several mutations demonstrably affecting the production of enzymes have been reported.[6] Evidence on the postulated gene-enzyme relationship is in most cases, however, still circumstantial; this is partly because of technical difficulties involved in the study of synthetic, or free-energy consuming reactions *in vitro*, and partly because of the insufficiency of biochemical information on those reactions which happen to be susceptible of genetic analysis.

As a corollary of the above hypothesis, each biosynthesis depends on the direct participation of a number of genes equal to the number of different, enzymatically catalyzed steps in the reaction chain. In attempting to account for the evolutionary development of such a reaction chain one meets in a clear form the problem of explaining macroevolutionary changes in terms of microevolutionary steps. The individual reactions making up the chain are of value to the organism only when considered collectively and in view of the ultimate product. Regarded individually, intermediate substances cannot, in general, be assumed to have physio-

logical significance, and the ability to produce them does not of itself confer a selective advantage. An example from *Neurospora* genetics will serve to illustrate this point. At the present time seven different genes are known to be concerned in the synthesis of arginine by the mold.[7] The inactivation of any one prevents the synthesis from taking place. On the basis of the above hypothesis, at least seven different catalyzed steps must occur in the synthesis. Several of the steps have been identified and controlling genes assigned to each. Two of the intermediates in the chain have been shown to be the amino acids ornithine and citrulline. Unlike arginine, neither of these substances is a general constituent of proteins. Aside from their function as precursors, they are apparently of no further use to the organism.

While the above example probably represents the general case, there are also well-known instances in which precursors serve independent functions. Thus, arginine, glycine and methionine are precursors of creatine in the rat,[8] but the synthesis goes through the non-functional intermediate, glycocyamine. On the other hand, acetylcholine may be synthesized from choline in one step.[9] In cases such as these, the problem is that of accounting for the synthesis of the precursors.

Since natural selection cannot preserve non-functional characters, the most obvious implication of the facts would seem to be that a stepwise evolution of biosyntheses, by the selection of a single gene mutation at a time, is impossible. It will be shown below that this is not a necessary conclusion, but that under special conditions the stepwise evolution of long-chain syntheses may occur. First, however, an alternative to stepwise evolution will be considered; that is, the origin of a new reaction chain through the chance combination of the necessary genes.

Although the probability of the origin of a useful character through the chance association of many genes may be small, it is never zero. Indeed, a consideration of the statistical consequences of the interaction of mutation, Mendelian inheritance, and natural selection has led Wright[10] to the conclusion that such chance associations may be of major importance in evolution. He has analyzed the evolutionary possibilities of various types of breeding structures and has shown that under certain conditions an extensive trial and error mechanism exists, whereby the species can test numerous combinations of non-adaptive genes. The breeding structure which most favors this type of evolution is that of a large population divided into many small, partially isolated groups. Within each group the cumulative effects of the accidents of sampling among the gametes are of major significance in determining gene frequencies, but the penalty of fixation of deleterious genes, ordinarily incurred under inbreeding, is avoided by exchange of migrants with other groups. The pressures of forward and reverse mutations, which between them determine an equilib-

rium frequency for non-adaptive genes in large, random-breeding populations, become of minor importance. As a consequence, a random drift of gene frequencies occurs. If, by chance, one group finds a particularly favorable combination of genes, a process of intergroup selection comes into play, whereby the favorable combination is spread to the population at large.

This model provides a means for the evolution of a new gene combination in spite of unfavorable mutation rates to active alleles and in the absence of selection of individual genes. It is thus favorable for the evolution of systems of individually non-adaptive, but collectively adaptive, genes. The effectiveness of the process would seem to be strongly dependent on the size of the gene combination required, however, decreasing approximately exponentially with increasing numbers of genes, other factors remaining constant. There would result a tendency toward the evolution of short reaction chains involving the recombination of molecular units already available. There is no doubt that a conservative tendency of this sort actually exists in nature. The wide variety of biologically important compounds built up on the pyrrole nucleus, to mention but one example, is a case in point.

The application of Wright's theory to the particular problem under consideration is limited by the fact that it operates only under biparental reproduction. It is probable that a large number of basic syntheses evolved prior to sexual reproduction. The universal distribution among living forms of certain classes of compounds—viz., the amino acids, nucleotides and probably the B vitamins—identifies them as essential ingredients of living matter. The synthesis of these substances must have evolved very early in geologic time, as a necessary condition for further progress, although loss of certain syntheses may have occurred in the later differentiation of some forms. It is therefore desirable to search for another solution of the problem applicable to compounds of this type, preferably one in which a minimum burden is placed on chance and a maximum one on directed evolutionary forces. It is thought that the following suggestion, while definitely a speculation, offers a possible solution along these lines.

In essence, the proposed hypothesis states that the evolution of the basic syntheses proceeded in a stepwise manner, involving one mutation at a time, but that the order of attainment of individual steps has been in the reverse direction from that in which the synthesis proceeds—i.e., the last step in the chain was the first to be acquired in the course of evolution, the penultimate step next, and so on. This process requires for its operation a special kind of chemical environment; namely, one in which end-products and potential intermediates are available. Postponing for the moment the question of how such an environment originated, consider the

operation of the proposed mechanism. The species is at the outset assumed to be heterotrophic for an essential organic molecule, A. It obtains the substance from an environment which contains, in addition to A, the substances B and C, capable of reacting in the presence of a catalyst (enzyme) to give a molecule of A. As a result of biological activity, the amount of available A is depleted to a point where it limits the further growth of the species. At this point, a marked selective advantage will be enjoyed by mutants which are able to carry out the reaction $B + C = A$. As the external supplies of A are further reduced, the mutant strain will gain a still greater selective advantage, until it eventually displaces the parent strain from the population. In the A-free environment a back mutation to the original stock will be lethal, so we have at the same time a theory of lethal genes. The majority of biochemical mutations in *Neurospora* are lethals of this type.

In time, B may become limiting for the species, necessitating its synthesis from other substances, D and E; the population will then shift to one characterized by the genotype $(D + E = B, \mathrm{B} + C = A)$. Given a sufficiently complex environment and a proportionately variable germ plasm, long reaction chains can be built up in this way. In the event that B and C become limiting more or less simultaneously, another possibility is opened. Under these circumstances symbiotic associations of the type $(F + G \neq C, D + E = B)(F + G = C, D + E \neq B)$ will have adaptive value.

This model is thus seen to have potentialities for the rapid evolution of long chain syntheses in response to changes in the environment. As has been pointed out by Oparin[11] the hypothesis of a complex chemical environment is a necessary corollary of the concept of the origin of life through chemical means. The essential point of the argument is that it is inconceivable that a self-reproducing unit of the order of complexity of a nucleoprotein could have originated by the chance combination of inorganic molecules. Rather, a period of evolution of organic substances of ever-increasing degree of complexity must have intervened before such an event became a practical, as distinguished from a mathematical, probability. Or, put in another way, any random process which can have produced a nucleoprotein must at the same time have led to the production of a profusion of simpler structures. Oparin has considered in some detail the possible modes of origin of organic compounds from inorganic material and cites a number of known reactions of this type, together with evidences of their large-scale occurrence on the earth in past geologic ages. He concludes that in the absence of living organisms to destroy them highly complex organic systems can have developed. The first self-duplicating nucleoprotein originated as a step in this process of chemical evolution. The origin of living matter by physicochemical means thus

presupposes the existence of a highly complex chemical environment.

To summarize, the hypothesis presented here suggests that the first living entity was a completely heterotropic unit, reproducing itself at the expense of prefabricated organic molecules in its environment. A depletion of the environment resulted until a point was reached where the supply of specific substrates limited further multiplication. By a process of mutation a means was eventually discovered for utilizing other available substances. With this event the evolution of biosyntheses began. The conditions necessary for the operation of the mechanism ceased to exist with the ultimate destruction of the organic environment. Further evolution was probably based on the chance combination of genes, resulting to a large extent in the development of short reaction chains utilizing substances whose synthesis had been previously acquired.

[1] Garrod, A. E., *Inborn Errors of Metabolism,* Oxford University Press (1923).

[2] Wright, S., *Biol. Symposia,* **6,** 337–355 (1942).

[3] Ephrussi, B., *Quart. Rev. Biol.,* **17,** 327–338 (1942).

[4] Lawrence, W. J. C., and Price, J. R., *Biol. Rev.,* **15,** 35–58 (1940).

[5] Horowitz, N. H., Bonner, David, Mitchell, H. K., Tatum, E. L., and Beadle, G. W., *Am. Nat.,* in press (1945).

[6] Summarized in Wright, S., *Physiol. Rev.,* **21,** 487–527 (1941).

[7] Srb, A., and Horowitz, N. H., *Jour. Biol. Chem.,* **154,** 129–139 (1944).

[8] Summarized in Schoenheimer, R., *The Dynamic State of Body Constituents,* Harvard University Press (1942).

[9] Lipmann, F., *Advances in Enzymology,* **1,** 99–162 (1941).

[10] Wright, S., *Bull. Am. Math. Soc.,* **48,** 223–246 (1942). Contains summary of earlier papers.

[11] Oparin, A. I., *The Origin of Life,* trans. by S. Morgulis, Macmillan, New York (1938).

*STRAIN SPECIFICITY AND PRODUCTION OF ANTIBIOTIC SUBSTANCES. V. STRAIN RESISTANCE OF BACTERIA TO ANTIBIOTIC SUBSTANCES, ESPECIALLY TO STREPTOMYCIN**

BY SELMAN A. WAKSMAN, H. CHRISTINE REILLY AND ALBERT SCHATZ

NEW JERSEY AGRICULTURAL EXPERIMENT STATION, RUTGERS UNIVERSITY, NEW BRUNSWICK, NEW JERSEY

Communicated April 17, 1945

Different strains of the same species of bacteria are found to vary greatly in their sensitivity to a given antibiotic substance. This phenomenon has an important bearing upon the utilization of the substance for chemotherapeutic purposes, where a knowledge of the sensitivity of the particular strain of a given organism responsible for a certain disease becomes of paramount importance.

It has been definitely established that when a culture of an organism is grown in the presence of a certain antibiotic substance, it becomes resistant to increasing concentrations of the substance; strains may thus be obtained that have become adapted to the action of the antibiotic agent and that require far greater concentrations for growth inhibition as compared with the parent or mother culture from which they were isolated. This has been shown to hold true for the action of lysozyme on *Micrococcus lysodeikticus*, and for various antibiotic and chemical agents against different bacteria. This phenomenon has been studied extensively in recent years as applied to the effect of penicillin on a number of bacteria, including *Staphylococcus aureus*,[1] pneumococci,[10] staphylococci, streptococci,[9] gonococci[2] and various others. Different strains of *S. aureus* have been shown[11] to vary in their susceptibility to penicillin in a range of less than 0.1 unit to greater than 1.0 unit.

Davies, Hinshelwood and Price[3] explained the phenomenon of adaptation of an organism to an antibacterial agent by one or more of the following mechanisms: (1) Adaptation occurs by natural selection from an initially heterogeneous population; this concept has lost, however, much support, since variations have been found to occur in strains derived initially from a single cell. (2) Adaptation occurs by actual modification of the individual cells; this may be due to the establishment in the cells of a mechanism alternative to that normally in use, or a quantitative modification of the existing mechanism occurs. (3) Adaptation is a change in some center of organization of the cell. In summarizing the various theories explaining the phenomenon of bacterial adaptation, Hinshelwood[7] concluded that when variations or adaptive changes occur, there is an actual modification of the character of the individual cells, although selection may be superimposed on this when modified and unmodified cells exist together. Demerec[4] concluded that resistance of *S. aureus* to penicillin originates through a mutation mechanism, and that penicillin acts as a selective agent to eliminate the non-resistant members of the bacterial population.

The adaptation of a certain bacterial strain to a given antibiotic substance does not necessarily signify that the strain is also resistant to another substance. A certain organism may become more resistant to penicillin by growth in gradually increasing concentrations of this antibiotic in an artificial culture medium, or more resistant strains of this organism may be isolated from patients treated with penicillin. Such strains do not necessarily show the same range of resistance to another antibiotic substance, like streptomycin. It was found,[15] for example, in a study of the variation of spore-forming bacteria, that different strains resistant to streptothricin are not at all resistant to clavacin or to fumigacin. Different strains of staphylococci were found[11] to show different

degrees of resistance to penicillin, tyrothricin and streptothricin, the action of one antibiotic not being necessarily paralleled by the action of another. When resistant strains are grown in media free from the antibiotic, a reversion may occur either gradually or suddenly and spontaneously; however, some of the resistant variants may be very stable.[7] The sensitivity of different strains of the same organism to a single antibiotic may also depend upon the morphological and physiological characteristics of the strains, as shown for the sensitivity of sporulating and non-sporulating strains of *Actinomyces griseus* to streptomycin.[14]

In the following investigations a study has been made of the sensitivity and adaptation of different strains of various gram-negative and gram-positive bacteria to streptomycin.[12] This antibiotic agent was selected because of its specific capacity to inhibit the growth of bacteria belonging to both gram-negative and gram-positive groups,[6, 8, 13] and because of its chemotherapeutic potentialities.[8]

TABLE 1

EFFECT OF STREPTOMYCIN UPON DIFFERENT STRAINS OF *E. coli* AND *E. communior*

E. coli		*E. communior*	
CULTURE NO.	D. U.[a]	CULTURE NO.	D. U.
ATCC 26	30	ATCC 130	30
ATCC 6522	30	ATCC 4163	30
ATCC 6880	10	ATCC 4351	30
ATCC 6881	30	ATCC 4352	30
ATCC 8677	100	ATCC 7011	30
ATCC 8739	30	Average	30
W 2	30		
W 4	30		
W 5	30		
Average	35		

[a] D. U. = dilution units per 1 mg. as determined by plate method.

Experimental.—Most of the cultures of bacteria used in these studies were obtained from the American Type Culture Collection (ATCC), though some were taken from our own collection (W), and some were received from special sources, as will be indicated. The streptomycin used in most of the experiments was a dry, fairly stable preparation produced in our own laboratory; it had an activity of 58 units per milligram.

In the first group of experiments, a number of pure cultures of bacteria comprising three gram-negative and two gram-positive species were tested for their sensitivity to streptomycin. The results presented in table 1 show that out of nine strains of *Escherichia coli* and five strains of *E. communior*, only one was very sensitive (100 units) to the antibiotic agent and one was very resistant (10 units); the remaining twelve strains were about equally sensitive, giving about 30 units as measured by the agar streak method.

Greater variation was obtained in the sensitivity to streptomycin of different strains of *Proteus vulgaris* and *S. aureus* tested by the same method, although the average values for these organisms as well were about the same as for *E. coli* (table 2). Out of seven cultures of *Pr. vulgaris*

TABLE 2

EFFECT OF STREPTOMYCIN UPON *S. aureus* AND *Pr. vulgaris*

Pr. vulgaris CULTURE NO.	D. U.	*S. aureus* CULTURE NO.	D. U.
ATCC 6898[a]	30	ATCC 152	25
ATCC 7829	75	ATCC 6518	25
ATCC 8259	50	ATCC 6538	45
ATCC 8427[a]	10	ATCC 8094	30
ATCC 9484	30	ATCC 8431	20
W 1[a]	30	W 1	30
W 2[a]	20	W 2	25
Average	35	W 4	75
		Merck Institute	30
		Average	34

[a] W 1 is same strain as 6898; W 2 is same as 8427.

used in these tests, one was very resistant (10 units) and one was rather sensitive (75 units), the remaining five cultures showing sensitivity to the particular streptomycin preparation ranging from 20 to 50 units, with an average of 35 units. The range in sensitivity of the nine *S. aureus* strains was from 20 to 75 units, with an average of 34 units.

The sensitivity of six strains of *Bacillus subtilis* to streptomycin is brought out in table 3. In the case of this organism, as well, there was considerable

TABLE 3

EFFECT OF STREPTOMYCIN UPON DIFFERENT STRAINS OF *B. subtilis*

DILUTION (PLATE) METHOD		DIFFUSION (CUP) METHOD, ZONE OF INHIBITION, MM.	
CULTURE	D. U.	DILUTION 1:5	DILUTION 1:25
ATCC 102	30	19.5	13.0
N. R. Smith 237	100	19.0	14.5
N. R. Smith 972	30	16.8	12.0
ATCC 6633[a]	250	26.3	20.5
ATCC 8473	100	23.0	17.5
W 9[a]	250	26.0	20.0
Average	109		

[a] ATCC 6633 is the same strain as W 9.

variation among the different strains. The tests with this organism were carried out by two methods, namely, the agar streak or plate dilution and the agar diffusion or cup methods. The degree of sensitivity of *B. subtilis* to streptomycin was found to be characteristic of the strain rather than of

the organism. Two strains, 102 and 972, were markedly resistant to streptomycin, as measured by both methods, the degree of resistance being of about the same order of magnitude as for the average strain of *E. coli*. Most of the strains of *B. subtilis* were, however, much more sensitive. The most sensitive strain was available in two different cultures, namely, ATCC 6633 and W 9; both showed the same degree of sensitivity by both methods of testing. Strain 237 produced by the cup method, in addition to the clear, easily readable zones, also secondary and even tertiary zones, which tended to confuse the readings, since only the clear zones were measured Finally, strain 8473 did not produce the same type of growth as did the other strains of *B. subtilis;* it formed rough surface colonies which gave the plates a peculiar appearance, although they did not interfere with the reading of the zone of inhibition. With the exception, therefore, of the last two cultures, the strains of *B. subtilis* varied in their sensitivity to streptomycin as much as 10:1.

TABLE 4

BACTERICIDAL EFFECT OF STREPTOMYCIN UPON *Pr. vulgaris*

INCUBATION WITH STREPTOMYCIN, HRS.	STREPTOMYCIN PER 1 ML. CULTURE			
	0	2	10	50
	MILLIONS OF CELLS PER 1 ML.			
0	1,290	1,290	1,290	1,290
6	2,150	1,010	665	100
24	1,470	110	30	2.3
48	745	1.2	0[a]	0[a]
124	360	0	0	0

[a] 1–4 colonies on a plate from 1:100,000 dilution.

In order to establish the manner in which strains of a given culture are capable of developing resistance against streptomycin, *Pr. vulgaris* was selected for more detailed studies. At first, the effect of the concentration of streptomycin upon its bactericidal action was determined by using a single strain (No. 2) of *Pr. vulgaris*. The results of a typical experiment are given in table 4. When 2.0 to 10 units of streptomycin were added to a culture of *Pr. vulgaris* all the cells were killed in 48 to 124 hours; when one unit was used, there was only a temporary inhibition of growth, which was later overcome. Cultures of *Pr. vulgaris*, treated with the larger amounts of streptomycin, were streaked, after 6 days' incubation, on nutrient agar plates, and the colonies produced from a few surviving cells were isolated and transferred to fresh media. A new strain was thus obtained possessing all the cultural and staining properties of the mother culture; however, it proved to be more resistant to streptomycin. This new strain was designated as R. Two original strains and the resistant

form were inoculated into several lots of fresh broth containing varying concentrations of streptomycin. The numbers of viable bacteria were determined after several periods of incubation at 37°C. The results reported in table 5 show emphatically that strain R, isolated from culture No. 2, became highly resistant to the action of streptomycin.

The sensitivity of the original culture of *Pr. vulgaris* and that of the resistant strain R was now measured against streptothricin, an antibiotic substance closely related to streptomycin. The results, which need not be reported here in detail, tended to demonstrate that strain R, made resistant to streptomycin, also became, though perhaps not to so great a degree, resistant to streptothricin, thus pointing to a certain similarity in the two substances. When, however, a totally different type of antibiotic agent was used, namely, clavacin, no difference could be found in the sensitivity of the streptomycin-resistant strain R as compared to that of

TABLE 5

BACTERICIDAL ACTION OF STREPTOMYCIN UPON *Pr. vulgaris* AND A FRESHLY ISOLATED RESISTANT STRAIN OF THIS ORGANISM

Number of Cells in Millions per 1 Ml.

STREPTOMYCIN UNITS PER 1 ML.	CULTURE W 1	CULTURE W 2	STRAIN R
	AT START		
	39.5	45.5	26.0
	AFTER 28 HRS.' INCUBATION		
0	375.0	615.0	560.0
1	0.01	0.65	190.0
5	0[a]	0[a,b]	9.0
25	0	0[a,b]	3.13

[a] 1 colony on a plate from 1:1000 dilution.
[b] 7–8 colonies on a plate from 1:10 dilution.

the mother culture No. 2. Thus, a strain of a bacterium that becomes resistant to a given antibiotic substance may also be resistant to a related compound, but not to a different type of antibiotic substance.

The results of a more detailed study of this phenomenon, using the mother culture and the resistant strain R of *Pr. vulgaris*, as well as two corresponding strains of *S. aureus*, are presented in table 6. With the first organism, the previous observations are confirmed. With *S. aureus*, however, the strain made more resistant to streptomycin was not rendered resistant to streptothricin. A possible explanation for this discrepancy may be found in the fact that the resistance of *S. aureus* to streptomycin was not increased sufficiently. The ratio of the activity of streptomycin against the parent strain of *Pr. vulgaris* to that against the resistant strain is about 12:1, whereas the ratio of the activities of streptothricin against these two strains is 6:1. With *S. aureus*, however, streptomycin was

only five times less active against the resistant strain than against the parent strain. Thus, if one assumes that these two agents act similarly against *Pr. vulgaris* and *S. aureus*, the resistance to streptomycin of the new strain of *S. aureus* is not sufficiently higher than that of the parent strain for differences in their resistances to streptothricin to become apparent.

The action of several antibiotic substances upon four cultures of *S. aureus* and their corresponding strains made resistant to streptomycin was now determined. These cultures were obtained from The Merck Institute of Therapeutic Research. The ratio of sensitivity to streptomycin of the resistant strains to that of the mother cultures varied from 750 to >2000. When tested against streptothricin, the corresponding ratios were 1 to 4, thus indicating only a very low degree of increase in resistance. When tested against clavacin and St, an antibiotic substance

TABLE 6

BACTERIOSTATIC ACTIVITY OF ANTIBIOTIC SUBSTANCES UPON STREPTOMYCIN-RESISTANT STRAINS OF TWO BACTERIA

	ACTIVITY, DILUTION UNITS PER MG.[a]				
	Pr. vulgaris			*S. aureus*	
ANTIBIOTIC AGENT	STRAIN 1	STRAIN 2[b]	RESISTANT STRAIN R	PARENT STRAIN	RESISTANT STRAIN
Streptomycin	250	100	8	250	50
Streptothricin	250	150	25	250	250
Clavacin	150	150	150	200	250

[a] The activity of the antibiotic agents was for streptomycin, 58 units; streptothricin, 500 units; and for the crystalline clavacin, 200 units; the actual results obtained for streptomycin were multiplied by 8.3, namely, the activity ratio of the streptothricin and streptomycin preparations, in order to make them comparable.

[b] Strain 2 is the parent culture from which the resistant strain R has been isolated.

produced by an aerobic spore-forming bacterium, no difference whatsoever in the sensitivity of the original cultures and streptomycin-sensitive strains could be detected.

Summary.—Different strains of the same bacterial species may vary greatly in their sensitivity to streptomycin. The ratio of sensitivity to a given preparation varied for *E. coli* from 100 to 10 units with an average of 35; for *Pr vulgaris*, 75 to 10 units (average 35); for *S. aureus*, 20 to 75 units (average 34); and for *B. subtilis*, 30 to 250 units (average 109).

A streptomycin-resistant strain of *Pr. vulgaris* showed also a certain degree of resistance to streptothricin but none at all to clavacin.

A strain of *S. aureus* that was made only slightly resistant to streptomycin showed no resistance to streptothricin. Several highly resistant strains of *S. aureus* showed no increase in resistance to clavacin and to an antibiotic substance isolated from a spore-forming soil bacterium, and only a trace of increased resistance to streptothricin.

* Journal Series Paper, New Jersey Agricultural Experiment Station, Rutgers University, Department of Microbiology.

[1] Abraham, E. P., Chain, E., Fletcher, C. M., Gardner, A. D., Heatley, N. G., Jennings, M. A., and Florey, H. W., *Lancet*, **241**, 177–189 (1941).

[2] Bahn, J. M., Ackerman, H., and Carpenter, C. M., *Proc. Soc. Exp. Biol. Med.*, **58**, 21–24 (1945).

[3] Davies, D. X., Hinshelwood, C. N., and Price, J. M., *Trans. Faraday Soc.*, **40**, 397–419 (1944).

[4] Demerec, M., *Proc. Nat. Acad. Sci.*, **31**, 16–24 (1945).

[5] Fleming, A., and Allison, V. D., *Brit. Jour. Exp. Path.*, **8**, 214 (1927).

[6] Heilman, F. R., *Proc. Staff Meet. Mayo Clinic*, **19**, 553–559 (1944).

[7] Hinshelwood, C. N., *Biol. Rev.*, **19**, 150–163 (1944).

[8] Jones, D., Metzger, H. J., Schatz, A., and Waksman, S. A., *Science*, **100**, 103–105 (1944).

[9] McKee, C. M., and Houck, C. L., *Proc. Soc. Exp. Biol. Med.*, **53**, 33–34 (1943).

[10] McKee, C. M., and Rake, G., *Ibid.*, **51**, 275–278 (1942).

[11] Neter, E., *Ibid.*, **58**, 126–128 (1945).

[12] Schatz, A., Bugie, E., and Waksman, S. A., *Ibid.*, **55**, 66–69 (1944).

[13] Schatz, A., and Waksman, S. A., *Ibid.*, **57**, 244–248 (1944).

[14] Schatz, A., and Waksman, S. A., *Proc. Nat. Acad. Sci.*, **31**, 129–137 (1945).

[15] Waksman, S. A., Bugie, E., and Reilly, H. C., *Bull. Torrey Bot. Club*, **71**, 107–221 (1944).

---

*STABILIZING INFLUENCE OF LIBERAL INTAKE OF VITAMIN A**

By H. C. Sherman and H. L. Campbell

Department of Chemistry, Columbia University, New York

Communicated May 12, 1945

Recently we have reported evidence that liberality of nutritional intake of vitamin A tends to postpone aging and to increase the length of life.[1]

It is now found further that offspring of the (rat) families having the higher vitamin A intakes grow not only somewhat more rapidly but also with distinctly lesser individual variability, thus indicating both a favorable and a stabilizing influence, of liberal vitamin A, upon the growth process.

Table 1 shows the mean weights at specified ages and the mean gain in weight from the 28th to the 56th days of age (end of infancy to early adolescence in the rat) for strictly parallel series of experimental animals on Diets 16, 360 and 361, which are otherwise alike but contain, respectively, 3, 6 and 12 International Units of vitamin A per gram of air-dry food; or 0.8, 1.6 and 3.2 I. U. per food calorie.

It will be seen that early growth is more rapid on the diets with 6 or 12 than on that with 3 I. U. per gram; while the growth was essentially alike on the two higher levels. This was equally true for each sex. Weights

TABLE 1

INFLUENCE OF THREE LEVELS OF DIETARY INTAKE OF VITAMIN A UPON GROWTH OF RATS: BODY WEIGHTS AT FIXED AGES, AND GAINS IN 5TH TO 8TH WEEKS OF AGE, IN GRAMS; WITH C. V. OF GAINS. (THE NUMBERS OF CASES ARE GIVEN IN PARENTHESES)

	MALES			FEMALES		
	ON DIET WITH 3 I. U./G.	ON DIET WITH 6 I. U./G.	ON DIET WITH 12 I. U./G.	ON DIET WITH 3 I. U./G.	ON DIET WITH 6 I. U./G.	ON DIET WITH 12 I. U./G.
Mean weight of offspring, continued on respective family diet, at age of						
28 days	(94) 43	(84) 44	(81) 45	(92) 43	(91) 42	(75) 43
56 days	(93) 98	(84) 114	(78) 115	(92) 87	(91) 101	(75) 102
84 days	(73) 160	(61) 184	(59) 184	(90) 129	(79) 151	(64) 149
100 days	(67) 189	(56) 215	(53) 213	(78) 146	(72) 166	(55) 163
120 days	(67) 215	(56) 244	(52) 242	(82) 163	(77) 173	(61) 172
150 days	(67) 250	(56) 273	(52) 271	(72) 177	(71) 187	(55) 184
180 days	(59) 273	(48) 293	(46) 290	(65) 186	(63) 193	(48) 190
210 days	(59) 288	(48) 308	(43) 304	(65) 193	(61) 203	(49) 199
240 days	(57) 300	(48) 319	(43) 314	(65) 199	(63) 206	(48) 205
270 days	(57) 308	(48) 325	(43) 321	(65) 203	(63) 215	(50) 212
300 days	(57) 313	(48) 328	(43) 326	(63) 207	(61) 217	(47) 217
365 days	(37) 320	(36) 334	(28) 334	(47) 219	(43) 231	(38) 229
400 days	(35) 322	(36) 334	(28) 336	(48) 221	(40) 231	(32) 233
500 days	(30) 323	(33) 327	(22) 323	(46) 229	(42) 236	(33) 235
600 days	(26) 309	(30) 308	(17) 306	(40) 220	(41) 230	(33) 231
700 days	(18) 291	(22) 294	(13) 292	(39) 213	(38) 222	(29) 223
730 days	(14) 296	(18) 292	(13) 287	(37) 212	(34) 220	(27) 221
Gains in weight, 28th to 56th days of age:						
Means in grams	(93) 54.2±0.89	(84) 69.6±0.81	(78) 70.4±0.70	(92) 44.0±0.90	(91) 59.2±0.53	(75) 59.2±0.55
Coefficient of variation	28.3	15.9	13.1	29.3	12.8	12.2

of pregnant females were not included in the averages for their respective age groups. Adult weights will be seen to be essentially alike for the 6- and 12-unit levels, and relatively little higher than for the corresponding animals on the intake level of 3 units per gram. The coefficients of variation of the body weights, as distinguished from gains, were small in all cases.

Much more striking are the facts brought out in the lower section of table 1 which show that in the period of rapid growth, between the 28th and 56th days of the life of these rats, the coefficient of variation of the individual data of the respective groups or series are, in both sexes, much larger for those on the 3 I. U. than for those on the 6- and 12-I. U. levels of vitamin A per gram of food. Yet rat families in our colony are thriving in the 58th generation on the diet containing even the lowest of these three levels.

We conclude that while 3 I. U. of vitamin A per gram of air-dry food (or 0.8 I. U. per food calorie) fully meets the requirements of adequacy, as the word is commonly understood, there is a somewhat higher and a much less variable rate of growth when the level of vitamin A intake is twice or four times higher.

This stabilizing effect appears to be a further advantageous influence of the same liberal levels of dietary vitamin A that have previously been shown[1] beneficial to adult vitality and length of life.

* Aided by grants from The Nutrition Foundation, Inc.

[1] Sherman, H. C., Campbell, H. L., Udiljak, M., and Yarmolinsky, H., these PROCEEDINGS, **31**, 107–109 (1945).

ILLUSTRATIONS AND SIMPLE ABSTRACT PROOF OF SYLOW'S THEOREM

BY G. A. MILLER

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF ILLINOIS

Communicated April 30, 1945

Sylow's theorem in the theory of finite groups consists of the following three parts: If the order of a group G is divisible by p^m, p being a prime number, but not by a higher power of p, then G contains at least one subgroup of order p^m, and if it contains more than one such subgroup all of these subgroups are conjugate under G and their number is of the form $1 + kp$. The first of these three parts can be proved independently of the other two parts. In fact, these two parts follow almost directly from the first of these three parts. Hence this part will receive most of our attention in what follows.

A Norwegian mathematician, L. Sylow (1832–1918), was the first to publish this general theorem. It appeared in the French language in a German mathematical journal called *Mathematische Annalen*, **5**, 584–594 (1872). This was about two years after the first treatise on group theory was published by a well-known French mathematician, C. Jordan (1838–1922). The title of this very influential work is *Traité des substitutions et des équations algèbriques* (1870). Various parts of this standard work of XVIII + 667 large pages could have been much simplified by the use of Sylow's theorem if it had then been known. Important steps toward its proof had been taken earlier. In particular, the noted French mathematician, A. L. Cauchy (1789–1857), published a proof of the fact that if the order of a group is divisible by a prime number then the group contains a subgroup whose order is equal to this number and hence the theorem has often been called the Cauchy-Sylow theorem. This special case had been stated without proof by E. Galois (1811–1832).

When G is an abelian group it can be proved very easily that Sylow's theorem applies to it by using the following obvious theorem relating to the order of the product of two commutative group operators. This order is the product of all the powers of the prime numbers which appear to a higher power in the order of one of these two operators then in the order of the other and the product of divisors of powers of prime numbers appearing to the same highest powers in the orders of the two given operators. The same theorem may be expressed by saying that if p^{α} is the highest power of p which divides the order of one of the two given operators but is not a divisor of the order of the other then it is a divisor of the order of their product but if p^{α} is the highest power of p which divides the orders of both of these operators and neither of them is divisible by a higher power of p then the order of their product is divisible by no higher power of p than p^{α}. The product of all such powers of prime numbers is the order of the product of the two given group operators.

From this theorem it follows directly that if none of the operators of the abelian group G has an order which is divisible by p then the order of G cannot be divisible by p and if the order of such an operator of G is divisible by p this operator is the product of an operator whose order is a power of p and an operator, which may be the identity, whose order is prime to p. Hence all the operators of G may be supposed to be so represented that some of them have orders which are powers of p while the rest of them have orders which are prime to p. The former will therefore generate a subgroup whose order is a power of p while the latter generate a subgroup whose order is prime to p and G is the direct product of these two subgroups. The order of the former subgroup is the highest power of p which divides the order of G and this is therefore the Sylow subgroup of order p^m contained in G.

It remains to consider the case when G is non-abelian, and it may first be noted that it may be assumed that G is one of the non-abelian groups of smallest order to which Sylow's theorem does not apply in case it does not apply to G. If G contains more than one invariant operator all of its invariant operators generate an invariant subgroup of G. Since both the order of this subgroup and the order of the corresponding quotient group are smaller than that of G it follows that Sylow's theorem applies to both of them and hence it also applies to G, which is contrary to the stated hypothesis. It therefore follows that it may be assumed that the identity is the only invariant operator of G.

All the operators of G can be arranged in sets of conjugates such that no two sets have any operator in common and that the identity alone constitutes one of these sets. Each of the operators of one of the sets is transformed into itself by all the operators of a subgroup of G and the number of the operators in the set is equal to the order of G divided by the order of a subgroup of G. Hence there results the following equation:

$$g = 1 + g_1 + g_2 + \ldots + g_n.$$

In this equation g represents the order of G and $g_1, g_2, \ldots, g_n$ represent the numbers of the operators in the given sets of conjugates. It should be noted that each of the numbers $g_1, g_2, \ldots, g_n$ usually exceeds unity and is equal to g divided by the order of a subgroup of G.

Since p divides g it cannot divide each of the numbers $g_1, g_2, \ldots, g_n$. Hence there is at least one of these numbers which is prime to p and each of the corresponding subgroups must have a common order which is divisible by p^m. These subgroups are found in G and hence G contains a subgroup of order p^m. To prove that all the subgroups of order p^m in G form a single set of conjugates it may be noted that a given one of these subgroups could not transform into itself another one of them since g is not divisible by a higher power of p than p^m. Therefore it results that if there were more than one set of such subgroups it would follow that if those of a set were transformed by one of their conjugates the number of the subgroups in the set would be of the form $1 + kp$ and if those of the same set were transformed by one of those in another set of conjugates this number would have to be of the form kp. Since this is a contradiction it has been proved that all these subgroups of order p^m are conjugate under G and that their number is of the form $1 + kp$.

The equation

$$g = 1 + g_1 + g_2 + \ldots + g_n$$

is fundamental. When $n = 1$, G is obviously the group of order 2. When

$n = 2$ G is either the group of order 3 or the symmetric group of order 6. When G is any abelian group n is clearly equal to $g - 1$ and hence it is only necessary to consider the cases when G is non-abelian. In all of these cases n is less than $g - 1$ and at least equal to the number of the distinct prime numbers which divide g. If it is equal to this number g cannot be divisible by the square of a prime number and hence it contains then an invariant subgroup whose order is the largest prime number which divides g. Hence the symmetric group of order 6 is the only group in which n is no larger than the number of the distinct prime numbers which divide g when G is a non-abelian group.

For every prime divisor p of g there is at least one of the numbers $g_1, g_2, \ldots, g_n$ which is prime to p and the sum of all such numbers when this sum is increased by unity is divisible by p. This includes the well-known theorem that every group of order p^m contains at least $p - 1$ invariant operators besides the identity, since the number of the operators in every set of conjugates of such a group is divisible by p whenever the set contains more than one operator. Hence the formula $g = 1 + g_1 + g_2 + \ldots + g_n$ which is useful in proving Sylow's theorem is also useful in the study of the prime power groups. A necessary and sufficient condition that G is an abelian group is that each of the numbers $g_1, g_2, \ldots, g_n$ is unity and hence n has then its maximal value.

When G is non-abelian the maximal value of n will clearly result when the number of the invariant operators of G is as large as possible and each of the non-invariant operators of G has only two conjugates under G provided that both of these conditions can be satisfied at the same time. This can be done in the present case. In fact, the largest number of the invariant operators of a non-abelian group can clearly not exceed the order of the group divided by 4 and when it is equal to this number the commutator subgroup of the group is of order 2 and hence each of the non-invariant operators of G has exactly two conjugates under G. In other words, when G is a non-abelian group the maximal value of n is $5g/8 - 1$ and the central of G is of order $g/4$ while its commutator subgroup is of order 2. The octic group is clearly the smallest group which satisfies these conditions.

A necessary and sufficient condition that the non-abelian group G contains a maximal number of sets of conjugate operators is that it contains three abelian subgroups of index 2. If it contains two such subgroups it also contains three such subgroups and its central is of index 4 under G while its commutator subgroup is of order 2. The theorem that a non-abelian group has three, one, or no abelian subgroups of index 2 is a special case of the theorem that a non-abelian group of order p^m has $p + 1$, one, or no abelian subgroup of index p, where p is a prime number. If a non-abelian group of order p^n contains more than one abelian subgroup of

index p the number of its sets of conjugate operators is $p^{m-2} + p^{m-1} - p^{m-3}$ and hence the maximal value of n in this case is $p^{m-3}(p^2 + p - 1) - 1$. This evidently reduces to $5g/8 - 1$ when $p = 2$ as was noted above.

ON THE NUMBER OF SOLUTIONS OF CERTAIN NON-HOMOGENEOUS TRINOMIAL EQUATIONS IN A FINITE FIELD

BY H. S. VANDIVER

DEPARTMENT OF PURE MATHEMATICS, UNIVERSITY OF TEXAS

Communicated May 7, 1945

The relation

$$au^m + bv^m + 1 = 0, \tag{1}$$

where a and b are given elements of a finite field F, with u and v to be determined in F, with $abuv \neq 0$, has been studied by Mitchell.[1] For the special case where F consists of residue classes with respect to a prime modulus, this equation has been studied by a number of writers.[2] All the methods used by these writers appear to depend on a certain symmetry arising from the fact that the exponents of u and v are the same.

However, in another paper[3] the writer obtained, based mainly on extensions of a method due to V. A. Lebesgue,[4] an expression for the number of roots of equation (1) modulo p which was independent of this symmetry. Hence, we shall apply this latter method to the consideration of

$$au^e + bv^f + w^g = 0 \tag{1a}$$

and

$$au^e + bv^f + 1 = 0. \tag{1b}$$

We shall find expressions for the number of solutions (u^e, v^f, w^g) of equation (1*a*) and the number of solutions (u^e, v^f) of equation (1*b*) in a finite field F, providing that $abuvw \neq 0$. If we obtain this number, then we may immediately find the number of solutions (u, v, w). Also we may confine ourselves to the case where e, f and g are divisors of $p^n - 1$, where the order of F is p^n, when p is prime, as the other cases depend on this.

Using Theorem II of my previous paper,[3] we determine N in the relation

$$N = \sum(1 - (ax_1 + bx_2 + x_3)^{p^n-1}), \tag{2}$$

when the summation extends over x_1, x_2, x_3, where x_1 ranges independently over all the distinct values such that $x_1^h = 1$, similarly x_2 over all roots of $x_2^i = 1$, and x_3 over all roots of $x_3^j = 1$, where

$$\frac{p^n - 1}{e} = h; \qquad \frac{p^n - 1}{f} = i; \qquad \frac{p^n - 1}{g} = j.$$

To reduce equation (2) we write

$$(ax_1 + bx_2 + x_3)^{p^n-1} = \sum_{k_1=0}^{p^n-1} (ax_1 + bx_2)^{k_1}(x_3)^{p^n-1-k_1}\binom{p^n - 1}{k_1}. \quad (2a)$$

Now

$$\binom{p^n - 1}{k_1} \equiv (-1)^{k_1}(\text{mod } p); \quad (2b)$$

and we may write, by further expansion,

$$(ax_1 + bx_2 + x_3)^{p^n-1} = \sum_{k_1=0}^{p^n-1} \sum_{r_1=0}^{k_1} (ax_1)^{r_1}(bx_2)^{k_1-r_1}(x_3)^{p^n-1-k_1}(-1)^{k_1}\binom{k_1}{r_1}$$

in the field. The summation of this with respect to x_3 leaves only terms involving $\binom{jk}{r_1}$, since $p^n - 1 - k_1 \equiv 0(\text{mod } j)$ must hold for such terms, and we set $jk = k_1$. Similarly, summation with respect to x_1 gives $hr = r_1$, and with respect to x_2 gives $jk - hr \equiv 0(\text{mod } i)$. This yields

$$\sum_{x_1,\, x_2,\, x_3} (ax_1 + bx_2 + x_3)^{p^n-1} = hij \sum_{k=0}^{g} \sum_{r} \binom{kj}{hr} a^{hr} b^{kj-hr}(-1)^{kj},$$

where r ranges over the values d in the set such that $jk - hd \equiv 0(\text{mod } i)$. Using this relation with equation (2), we have

THEOREM I. *If N is the number of distinct solutions (u^e, v^f, w^g) of the equation*

$$au^e + bv^f + w^g = 0$$

in a finite field F of order p^n, which contains a, b and c; $abuvw \neq 0$, then

$$N = -hij \sum_{k=1}^{g} \sum_{r} \binom{kj}{hr} a^{hr} b^{kj-hr} (-1)^{kj} \quad (3)$$

in F. Here e, f and g each divide $p^n - 1$;

$$h = \frac{p^n - 1}{e}; \qquad i = \frac{p^n - 1}{f}; \qquad j = \frac{p^n - 1}{g};$$

and r ranges over the values d in the set

$$0, 1 \ldots \left[\frac{jk}{h}\right]$$

such that $jk - hd \equiv 0$ (mod i).

In the same way we may consider the equation

$$au^e + bv^f + 1 = 0, abuv \neq 0, \tag{4}$$

and obtain by the same method, for the number of its solutions (u^e, v^f) in $F(p^n)$

$$-hi\sum_{k=1}^{f} \sum_{r=0}^{c} \binom{ki}{hr} a^{hr}b^{p^n-1-ki} (-1)^{ki}, \tag{5}$$

where

$$c = \left[\frac{ki}{h}\right].$$

We may then state the following:

THEOREM II. *If N is the number of distinct sets of solutions (u^e, v^f) in a finite field $F(p^n)$ of*

$$au^e + bv^f + 1 = 0$$

where $abuv \neq 0$, then N has the form (5), in the field.

The number of solutions of equation (4) is at most $p - 1$, for $n = 1$, so that we have the

THEOREM III. *The number of sets of distinct integral solutions (u^e, v^f) of*

$$au^e + bv^f + 1 \equiv 0 (\text{mod } p), \tag{6}$$

where a and b are integers, p is prime $(abuv, p) = 1$,

$$h = \frac{p-1}{e}, \qquad i = \frac{p-1}{f},$$

is the least residue $\geqq 0$ of

$$-hi \sum_{k=1}^{f} \sum_{r=0}^{c} \binom{ki}{hr} a^{hr}b^{p-1-ki} (-1)^{ki} \tag{7}$$

modulo p, where

$$c = \left[\frac{ki}{h}\right].$$

In the proof of Theorem I we could also have written the right-hand member of equation (2a) in the form

$$\sum_{k_1=0}^{p^n-1} (ax_1 + x_3)^{k_1}\, bx_2^{p^n-1-k_1}\binom{p^n-1}{k_1},$$

as well as

$$\sum_{k_1=0}^{p^n-1} (bx_2 + x_3)^{k_1}\, ax_1^{p^n-1-k_1}\binom{p^n-1}{k_1}.$$

This shows that if we call the expression (7), $A(h, i, j)$, then this equals in the field, each one of the expressions obtained by interchanging the letters h, i, j in the proof of Theorem I, namely: $A(i, h, j)$, $A(i, j, h)$, $A(j, i, h)$, $A(h, j, i)$ and $A(j, h, i)$; and these are all equal to N in the field. This fact does not appear easy to prove by any direct method.

As an application, the number, N, of sets of distinct integral solutions (x^4, y^2) of

$$ax^4 + by^2 + 1 \equiv 0 (\bmod p)$$

where a and b are integers, p is prime $(axby, p) = 1$, and $h = \frac{p-1}{4}$, is seen by Theorem III to be equal to the least integral residue modulo $p \geqq 0$ of

$$-\frac{p-1}{4} \cdot \frac{p-1}{2} \sum_{k=1}^{2} \sum_{r=0}^{2k} \binom{2kh}{hr} a^{hr} b^{p-1-2kh} (-1)^{2kh}.$$

Modulo p, this reduces, after using relation (2b), to

$$-{}^1/_8 \left(b^{2h} + \binom{2h}{h} a^h b^{2h} + a^{2h} b^{2h} + 1 + (-1)^h a^h + a^{2h} + (-1)^h a^{3h} + 1\right). \tag{8}$$

As is known p can be expressed uniquely in the form

$$p = (|r|)^2 + (|s|)^2 \tag{8a}$$

with r and s of different parity. Let us select s to be positive and even and choose the sign of r so that r has the form $4n - 1$ if p has the form $8n + 1$, but so that r has the form $4n + 1$ if p is of the form $8n + 5$. It is also known[5] that in this case

$$(-1)^{h+1} 2r \equiv \binom{2h}{h} (\bmod p). \tag{8b}$$

Let g be a primitive root of p defined by the congruence

$$rg^h + s \equiv 0 (\bmod p). \tag{8c}$$

Such a g will always exist; and a can be expressed as g^t;

$$t = 0, 1, 2, 3.$$

Therefore, using relations (8b) and (8c), we find

$$N \equiv -{}^1/_8 (b^{2h} + (-1)^{h+1}\, 2rg^{th} b^{2h} + g^{2th} b^{2h} + 5), \tag{9}$$

if

$$(-1)^h g^{th} \equiv 1 (\bmod p); \tag{10}$$

and

$$N \equiv -{}^1/_8(b^{2h} + (-1)^{h+1}\,2rg^{th}b^{2h} + g^{2th}b^{2h} + 1), \tag{11}$$

if

$$(-1)^h g^{th} \not\equiv 1 (\text{mod } p). \tag{12}$$

Let either $t = 0$, and h be even; or $t = 2$, and h be odd. Then relations (9) and (10) apply, and congruence (9) reduces to

$$N_1 \equiv -{}^1/_8\left(2\left(\frac{b}{p}\right) - 2r\left(\frac{b}{p}\right) + 5\right), \tag{13}$$

where $\left(\frac{b}{p}\right)$ is the Legendre symbol. Now let either $t = 0$, and h be odd; or $t = 2$, and h be even. Then congruences (11) and (12) apply, and relation (11) reduces to

$$N_2 \equiv -{}^1/_8\left(2\left(\frac{b}{p}\right) + 2r\left(\frac{b}{p}\right) + 1\right), \tag{14}$$

while if $t = 1$ or $t = 3$, then relation (11) applies and can be written, using relation (8c)

$$N_3 \equiv -{}^1/_8\left((-1)^h 2s\left(\frac{g}{p}\right)^{(t-1)/2}\left(\frac{b}{p}\right) + 1\right). \tag{15}$$

These formulae determine N as a least positive or zero residue modulo p. However, it is possible when $p > 13$ to exhibit N as an integer without any reduction with respect to the modulus p, as shall now be shown.

Designate by α_d, $d = 1, 2, 3$, respectively, the number $-8N$ in relations (13), (14) and (15). Then

$$|\alpha_1| \leqq 7 + 2|r|,\ |\alpha_2| \leqq 3 + 2|r|,\ |\alpha_3| \leqq 1 + 2s$$

and from relation (8a)

$$|r| < \sqrt{p},\ s < \sqrt{p}$$

so

$$|\alpha_d| \leqq 7 + 2\sqrt{p}.$$

When $p > 13$, then it is easy to show that

$$7 + 2\sqrt{p} < p, \tag{15a}$$

and

$$|-\alpha_d| + 7p < 8p. \tag{16}$$

Now since $N \equiv -{}^1/_8\alpha_d(\text{mod } p)$, or $8N \equiv -\alpha_d + zp$, it follows that $0 \equiv$

$-\alpha_d + zp \pmod 8$. Since p and 8 are relatively prime, this linear congruence in z has just one non-negative integral solution $\theta < 8$.

$$8N \equiv -\alpha_d + \theta p \pmod p. \tag{17}$$

Since $|-\alpha_d + \theta p| < 8p$, and N is the least non-negative integral solution of equation (17), therefore N equals the lesser non-negative one of the following two numbers:

$$-{}^1/_8(\alpha_d - \theta p); \qquad -{}^1/_8(\alpha_d - (8 + \theta)p).$$

The same reasoning will apply to the case $p = 13$ if $\alpha_d \not\equiv 0 \pmod{13}$.

THEOREM IV. *The number, N, of sets of distinct integral solutions (x^4, y^2) of*

$$ax^4 + by^2 + 1 \equiv 0 \pmod p \tag{18}$$

where a and b are integers, p is a prime > 13 of the form $4h + 1$, and $(axby, p) = 1$, is equal to the lesser non-negative one of the following two numbers:

$$-{}^1/_8(\alpha_d - \theta p),$$
$$-{}^1/_8(\alpha_d - (8 + \theta)p),$$

where θ is the least non-negative integral solution z of the congruence

$$0 \equiv -\alpha_d + zp \pmod 8,$$

and where α_d, $d = 1, 2, 3$, is taken to be equal, respectively, to $-8N_1$, $-8N_2$, $-8N_3$ in congruences (13), (14), (15).

COROLLARY. *The congruence (18) of Theorem IV for $p > 13$ always has solutions.*

This follows from the inequality (15a) and the fact that α_d is odd and hence $\neq 0$.

[1] *Ann. Math.*, II, **18**, 120–131 (1917).

[2] Cf. the references given by Mitchell in his article above cited, as well as Dickson, *Am. Jour. Math.*, **57**, 391, 463 (1935); *Trans. Amer. Math. Soc.*, **37**, 363 (1935).

[3] These PROCEEDINGS, **30**, 362–367 (1944).

[4] *Jour. de Math.*, I, **2**, 260–266 (1837). The writer was in error in ascribing this method, as applied to trinomial congruences involving rational integers, to Hurwitz. Lebesgue in the reference just cited went much further than Hurwitz, obtaining a result which is closely related to Theorem III of the article referred to in the third footnote above, page 367.

[5] Bachmann, *Die Lehre von der Kreisteilung*, 133, 135, 136 (1872).

PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 31 July 15, 1945 Number 7

A CONTRIBUTION TO THE CYTOLOGY OF THE AUSTRALIAN-SOUTH PACIFIC SPECIES OF NICOTIANA*

BY HELEN-MAR WHEELER

DEPARTMENT OF BOTANY, UNIVERSITY OF CALIFORNIA

Communicated June 4, 1945

Although most of the fifty-eight species of the genus *Nicotiana* (Goodspeed[1]) are restricted to North and South America and adjacent islands, fifteen are peculiar to Australia and the South Seas. Of these, thirteen are limited to Australia, one, *N. Debneyi*, to Australia, New Caledonia and Lord Howe Island, and one, *N. fragrans*, occurs on the Isle of Pines near New Caledonia and eastward in a few scattered localities to the Marquesa Islands in the mid-South Pacific. As was recognized by Goodspeed,[2] the fifteen Australian-South Pacific species form a "genetic group" which he[1] has now designated as the section SUAVEOLENTES of the subgenus PETUNIOIDES. Its nearest relatives today are South American species of the ALATAE, ACUMINATAE and NOCTIFLORAE sections of that subgenus. The readiness with which morphological elements that give individuality to each of those three sections can be recognized in members of the SUAVEOLENTES suggests that all four sections may have been derived from a common source.

Detailed descriptions of species in the SUAVEOLENTES were given by the author[3] in 1935. The general pattern is that of an herb with conspicuous basal rosette or cushion of leaves and short, slender, forking stems terminating in long, spreading, loosely branching inflorescences. *N. suaveolens*, a variable species of medium flower size (25–45 mm. long) and obovate cauline leaf approximates the morphological mean. From it, *N. fragrans*, *N. Debneyi* and *N. Benthamiana* are the most extreme departures although only the last-named species is separated by any considerable hiatus unbridged by other species. *N. fragrans* has certain of the characters of *N. suaveolens* and also some of a long-flowered species, *N. megalosiphon*. *N. Debneyi* is approached by the small-flowered species, *N. exigua*, and in leaf and indument by *N. occidentalis*. In addition, *N. maritima* may be placed close to *N. velutina*, *N. suaveolens* and *N. Gossei*; *N. Gossei* to *N. excelsior;*

N. Goodspeedii to *N. exigua* and *N. rotundifolia; N. rotundifolia* to *N. occidentalis.* In other words, the section SUAVEOLENTES is an herbaceous group, the individual species of which appear to be complexly interrelated and in general separated by distinctions of only small magnitude. As a contribution to the cytology of this geographically isolated section and its interpretation, this study presents data on chromosome number and morphology in thirteen species and evidence bearing upon chromosome homology in twelve.

Chromosome Number.—According to Goodspeed[4] twenty-eight of the American species of *Nicotiana* are characterized by a somatic chromosome number of 12 pairs, nine species have 24 pairs, three have 9 pairs, two have 10 pairs. Among the Australian-South Pacific species there are no 9-, 10- or 12-paired ones but, as previously reported by Goodspeed and Clausen,[5] East (from Goodspeed),[6] Goodspeed (from Wheeler),[2,7] Kostoff, Dogadkina and Tichonova,[8] Wheeler,[3] Kostoff,[9,17] Kostoff (from Clausen),[9] there were four species which possess 16, three 18, two 20 and one each 22, 24 or 32† pairs. The author has now determined that two species, *N. Benthamiana* and *N. excelsior*, stated by Kostoff[9,17] to be 18-paired are 19 paired in her material, and in addition reports 21 pairs for *N. occidentalis* and 24 pairs for *N. fragrans*, species for which chromosome number had not previously been determined. Thus, as shown in table 1, the series begins with 16 pairs, closes with 24 and lacks only 17- and 23-paired species of being continuous. It is composed of four species with 16 pairs (one containing 32-paired races†), one with 18, two with 19, two with 20, one with 21, one with 22 and two with 24. The species which is selected as the morphological mean, *N. suaveolens*, has 16 pairs. *N. fragrans* and *N. Debneyi*, outstanding variants in the group, have 24 pairs. The species of greatest morphological isolation, *N. Benthamiana*, has 19 pairs. Identity or similarity in chromosome number may or may not parallel similarity in external morphology. In the case of three of the 16-paired species, of an 18-paired and one of the 19-paired (*N. excelsior*), and of a 21- and 22-paired, it does. By contrast, the lack of correlation is striking in the two 20-paired and the two 24-paired species.

Chromosome Morphology.—The only comprehensive work on karyotypes in the American species of *Nicotiana* is that of Goodspeed.[7,4] Those species have an average chromosome length of from 2.2 μ in the genom of shortest chromosomes to at least 5.5 μ in the genom of longest chromosomes. Three categories of chromosomes are recognized: median ("*x*") chromosomes in which the centromere position is median, submedian ("*sm*") chromosomes where it is nearer the middle than the end, and subterminal ("*st*") chromosomes where it is as near to the end as to the middle or nearer. Some species show only *m* or *sm* chromosomes, most have more *m* and *sm* than *st* chromosomes, and four (two each of sections

NOCTIFLORAE and ALATAE in the subgenus PETUNIOIDES) have exclusively *st* chromosomes. Satellites occur on at least one pair in each species, on two pairs in some, and rarely, in certain 24-paired species, on three pairs. In *N. Langsdorffii* there is an exceptionally large satellite. Constrictions not associated with centromere or satellite occur in some species.

For the Australian-South Pacific species partial analyses of the somatic complements of five species were reported by the author[12] and referred to by Goodspeed.[7] This earlier work was based upon paraffin sections of root tips. The complete and detailed analyses of chromosome morphology in table 1 of thirteen species of the section SUAVEOLENTES are products of recent studies in aceto-carmine smears of first microspore mitoses. For comparison with the rest of the genus the chromosomes are classified as to centromere position according to the terminology adopted by Goodspeed.

The length of the longest chromosome in its fully expanded condition in the SUAVEOLENTES is approximately 5 μ while the shortest is less than 2 μ. From species to species average chromosome length tends to decrease as number of chromosomes increases. Within a particular complement the longest chromosome usually is approximately twice the length of the shortest although in *N. Benthamiana*, *N. Debneyi* and *N. fragrans* the distinction is much less; the transition from longest to shortest is gradual.

As shown in table 1 there is no Australian-South Pacific species which possesses either all *m* or *sm* chromosomes or all *st* chromosomes. The lowest per cent of *st* (25%) occurs in the 16-paired species, *N. suaveolens*, *N. maritima* and *N. velutina*, the highest (79%) in the 24-paired species, *N. Debneyi*. In the main, as number of chromosomes in the complement increases there is progressive increase in number of *st* at the expense of *m* and *sm* chromosomes, but the relation is not a perfect one. *N. fragrans*, a 24-paired species, shows only 67 per cent of *st* and *N. Benthamiana*, a 19-paired, 74 per cent. Throughout, a proximal satellite occurs either on one pair of *st* chromosomes or on a pair which is almost on the border line between an *sm* and an *st* centromere position. In *N. fragrans* two pairs with proximal satellites apparently occur. Another satellite, in some species on the shorter arm of an *sm* chromosome and in others on an *m* chromosome, has been demonstrated in eleven species. It may also occur in *N. occidentalis* but the evidence is not conclusive. In *N. Benthamiana* a large segment separated by a distinct thread and peculiar to an *sm* chromosome is conspicuous. It is a question whether it has replaced the satellite mentioned and, if so, whether it is characteristic of all races of the species. In appearance it approximates the large satellite found in *N. Langsdorffii* of the ALATAE section (cf. Goodspeed[7]). Possibly more than two satellites will be found in *N. Debneyi*; there are six nucleoli in this species. On one to four *m* or *sm* chromosomes of all species except *N. Benthamiana* an additional slight constriction has been noted. When

most clearly shown, the chromosome arm containing it appears segmented, although in other cases a flexure may be the only indication.

In general it may be said of the karyotypes of the Australian-South Pacific species of *Nicotiana* that: (1) They show no morphological feature not also present elsewhere in the genus and particularly in those PETUNIOIDES

TABLE 1

CHROMOSOME MORPHOLOGY

SPECIES	TOTAL PAIRS	*m* PAIRS	*sm* PAIRS	*st* PAIRS
N. suaveolens Lehm.	16	8 (1s)	4 (1s)	4
N. maritima Wheel.	16	8 (1s)	4 (1s)	4
N. velutina Wheel.	16	8 (1s)	4	4 (1s)
N. exigua Wheel.	16	6 (1s)	5	5 (1s)
N. Gossei Domin	18	5	4 (1s)	9 (1s)
N. excelsior Black	19	5	4 (1s)	10 (1s)
N. Benthamiana Domin	19	1	4 (1s?)	14 (1s)
N. megalosiphon Heurck and Muell. Arg.	20	2	4 (1s)	14 (1s)
N. Goodspeedii Wheel.	20	6 (1s)	4 (1s)	10
N. occidentalis Wheel.	21	4	3	14 (1s)
N. rotundifolia Lindl.	22	4 (1s)	3	15 (1s)
N. Debneyi Domin	24	4 (1s)	1	19 (1s)
N. fragrans Hook.	24	2 (1s)	6	16 (2s?)

m, centromere median; *sm*, submedian; *st*, subterminal. 1s, a satellite on one pair

sections of closest resemblance (cf. Goodspeed[4]). (2) While no single karyotype pattern obtains consistently throughout the section save that all complements have *m*, *sm* and *st* chromosomes, nevertheless there is unity, for there is no complement that cannot be related morphologically in some degree to that of every other species and, usually, related closely to that of one or more species. In this respect, chromosome morphology sometimes gives better indication of degree of relationship than chromosome number alone. A case in point is the closer relationship of *N. Gossei* (18 pairs) to *N. excelsior* (19) than of either species to *N. Benthamiana* (19). (3) If one wishes to consider the section that in the light of Lewitsky's[13] hypothesis that complements of equal-sized, equal-armed chromosomes usually are the more primitive (cf. also Sarana,[14] Goodspeed[15]), phyletically the SUAVEOLENTES appear to be one of the more specialized groups in the genus and to have proceeded a long distance from the condition described by Goodspeed for the PANICULATAE section of the subgenus RUSTICA. However, no individual species of the SUAVEOLENTES has a complement consisting entirely of unequal-armed chromosomes. Hence, according to Lewitsky's hypothesis, no SUAVEOLENTES species has reached the degree of arm length specialization possible for its chromosome number which *N. longiflora* of the ALATAE section as described by Hollingshead[16] has attained.

Chromosome Homology.—Chromosome association to form bivalents or higher valencies at the first meiotic metaphase is considered here to be evidence of homology, i.e., evidence of a high degree of genic and structural similarity of entire chromosomes or of sufficiently large segments thereof to favor chiasmata formation. Non-homologous association may occur but its incidence is assumed to be negligible. In what follows, therefore, homology is discussed in terms of pairing.

Chiefly Kostoff[17,9] and Goodspeed[7,15] have studied the cytology of thirty-eight F_1 hybrids between Australian-South Pacific and American species. Evidence of a high degree of genic or structural distinction between the species involved in any given hybrid is indicated by the fact that in every case, except five involving a 32-paired race, pairing at MI is largely lacking. This also shows that in the complements of the SUAVEOLENTES species concerned, *N. suaveolens*, *N. maritima*, *N. exigua*, *N. Gossei*, *N. megalosiphon* and *N. Debneyi*, there is either little or no autosyndetic pairing.

Six interspecific F_1 hybrids between species of the SUAVEOLENTES were early examined cytologically by the author[12] but the first complete MI analyses of such intrasectional hybrids were published by Kostoff.[17,9] Exclusive of those in which a 32-paired race was a parent, he reported on the extent and quality of pairing in thirteen combinations. More recently the author has examined ten of the same and also investigated an additional sixteen hybrids. As far as it is possible to compare the data there is similarity between Kostoff's results and those listed in tables 2 and 3.

In the following tables when the parents differ in chromosome number the species of the lower number is placed first, irrespective of the direction in which the cross was made. Only pollen mother cells were studied and the amount and nature of chromosome association is reported on the basis of the first ten MI configurations which were completely analyzable in side view. Ten complete analyses were considered adequate for the purpose since studies of several times that number of pollen mother cells in a series of representative hybrids did not significantly alter the original conclusions. For example, where more than one "most frequent association" is recorded for ten counts, a sharply defined mode was not obtained with additional counts.

The above tables indicate that trivalents were seen in all but three of the twenty-six hybrids investigated. In more than half of them the most frequently observed association involved a trivalent and in two hybrids every pollen mother cell analyzed contained at least one valency higher than two. The occurrence of quadrivalents is less than that of trivalents, but one or more occurred within ten counts in thirteen hybrids. In a subsequent publication the incidence of all the types of chromosome association will be discussed in detail. Here only a summary of the some-

what complex situation concerned has been attempted. Column 2 gives the range of bivalents observed, column 3 the range in number of pairs calculated on the basis that one trivalent is equivalent to one pair and one quadrivalent to two pairs (cf. Goodspeed[15]), and column 4 the most frequent number of pairs in terms of a similar calculation. In column 4 the derivation of the number of pairs is shown by the number of trivalents in parentheses. For example, in F_1 *N. maritima* × *N. velutina* the "most

TABLE 2

CHROMOSOME ASSOCIATIONS AT MI IN F_1 HYBRIDS WHICH INVOLVE 16-PAIRED SPECIES

PARENTAL SPECIES	RANGE OF OBSERVED BIVALENTS	RANGE OF ASSOCIATIONS EXPRESSED AS BIVALENTS[a]	MOST FREQUENT ASSOCIATIONS EXPRESSED AS BIVALENTS[a]
16 × 16			
[K]*N. maritima—N. suaveolens*	16	16	16
[K]*N. maritima—N. velutina*	7–14	12–15 ($1–5_{III}$[b])	14 (2_{III})
[K]*N. velutina—N. exigua*	9–14	12–15 ($0–2_{III}1_{IV}$)	14 (1_{III})
[K]*N. velutina—N. suaveolens*	11–15	13–15 ($0–2_{III}$[b])	14
16 × 18			
[K]*N. maritima—N. Gossei*	12–16	14–16 ($0–1_{III}1_{IV}$)	15 (1_{III})
N. velutina—N. Gossei	12–15	13–16 ($0–1_{III}1_{IV}$)	15
[K]*N. suaveolens—N. Gossei*	13–15	13–15 ($0–1_{IV}$)	15 (1_{III})
N. exigua—N. Gossei	10–14	13–15 ($1–4_{III}$[b])	14 (2_{III})
16 × 19			
N. suaveolens—N. Benthamiana	7–14	8–14 ($0–1_{III}$)	8; 10; 11; 12 ($0–1_{III}$)
16 × 20			
N. suaveolens—N. Goodspeedii	14–17	16–17 ($0–2_{III}$[b])	16 (1_{III}); 17
N. exigua—N. Goodspeedii	11–16	14–16 ($0–1_{III}1_{IV}$)	14 (2_{III}); 15 (1_{III})
[K]*N. suaveolens—N. megalosiphon*	9–15	10–15 ($0–1_{III}$)	13 ($0–1_{III}$); 15 (1_{III})
[K]*N. exigua—N. megalosiphon*	10–14	10–14 ($0–1_{III}$)	13; 14
N. velutina—N. megalosiphon	7–13	8–14 ($0–1_{III}$)	12 (1_{III})
16 × 22			
N. velutina—N. rotundifolia	13–16	13–16	15; 16
16 × 24			
[K]*N. maritima—N. Debneyi*	11–16	14–16 ($0–3_{III}$)	16 (3_{III})

[K], Kostoff also reports this cross.

[a] Univalents omitted.

[b] Quadrivalents occur in this hybrid but do not increase the range.

frequent association" is stated to be "14_{II} (2_{III})." This indicates that pollen mother cells containing $12_{II} + 2_{III} + 2_{I}$ were most often found. In column 3, although the range in number of pairs also is a product of translation of multivalents into pairs, the figures in parentheses express only the range of occurrence of trivalents and quadrivalents or combination of the two.

Exclusive of hybrids in which *N. Benthamiana*, *N. megalosiphon* and *N. occidentalis* are parents, it can be seen in tables 2 and 3 that the number of pairs is or approaches 16 if both parental species are 16-paired, or ap-

proximates that of the parental species possessing the smaller chromosome number if the two parents differ from each other in chromosome number (i.e., "Drosera scheme" association). Thus nine of the twelve species of the SUAVEOLENTES investigated exhibit a high degree of homology and thereby demonstrate concretely common origin even in instances of considerable divergence of chromosome number and chromosome morphology.

The hybrids which involve *N. Benthamiana* show the greatest range of pairs (6 to 15), the lowest mode when a mode occurs (11 pairs), and the

TABLE 3

CHROMOSOMIC ASSOCIATIONS AT MI IN F_1 HYBRIDS WHICH INVOLVE 18- TO 24-PAIRED SPECIES ONLY

PARENTAL SPECIES	RANGE OF OBSERVED BIVALENTS	RANGE OF ASSOCIATIONS EXPRESSED AS BIVALENTS[a]	MOST FREQUENT ASSOCIATIONS EXPRESSED AS BIVALENTS[a]
18 × 19			
N. Gossei—*N. excelsior*	16–18	17–18 (0–1_{III})	17 (1_{III}); 18
N. Gossei—*N. Benthamiana*	9–14	9–15 (0–1_{III})	11 (1_{III})
18 × 20			
[K]*N. Gossei*—*N. megalosiphon*	11–16	13–17 (0–$1_{III}1_{IV}$)	15 (1_{III})
18 × 21			
N. Gossei—*N. occidentalis*	11–14	13–15 (0–2_{III}[b])	14
19 × 20			
N. Benthamiana—*N. megalosiphon*	6–14	6–15 (0–1_{III})	6; 10; 14 (0–1_{III}).
19 × 24			
N. Benthamiana—*N. Debneyi*	9–16	9–17 (0–$1_{III}1_{IV}$)	11
20 × 20			
N. Goodspeedii—*N. megalosiphon*	12–16	12–16	13; 15
20 × 22			
N. Goodspeedii—*N. rotundifolia*	17–20	17–20 (0–2_{III})	19 (1–2_{III})
20 × 24			
N. Goodspeedii—*N. Debneyi*	15–19	16–20 (0–2_{III}[b])	19 (1_{III})
N. megalosiphon—*N. Debneyi*	9–15	9–16 (0–1_{III})	11; 14

[K], Kostoff also reports this cross.
[a] Univalents omitted.
[b] Quadrivalents occur in this hybrid but do not increase the range.

least tendency to form a mode. Those of *N. megalosiphon* and the one *N. occidentalis* hybrid may exhibit some or all of these characteristics, but to lesser degree. However, in hybrids with any of these three species as one of the parents, the range of calculated pairs reaches or approaches 16. Thus extent of pairing here also, indicates both considerable homology between these species and other Australian species and, possibly, some significance in the number 16.

The consistent formation of one or a few trivalents and an occasional quadrivalent in many Australian hybrids may be explained on the basis of translocation or duplication, as Kostoff[17] has suggested. Aside from 32-paired races of *N. suaveolens*, a special problem, Kostoff[17] submits certain

evidence of one large duplication in *N. megalosiphon*. The author's counts of F_1 *N. Goodspeedii* × *N. suaveolens* suggest a similar possibility in *N. Goodspeedii*. However both instances require further investigation before it can be said that this is true of either *N. megalosiphon* or *N. Goodspeedii* as a species. The most probable explanation of multivalency in many Australian hybrids is that translocation is involved.

Conclusion.—The somewhat specialized character of the Australian-South Pacific species, the general similarity in external morphology from species to species, and the apparent intricacy of interrelationships among them, are reflected in their cytology. Closer relationship between certain species and a more distant one in the case of others is indicated sometimes by chromosome number or morphology, sometimes only by chromosome behavior in hybrids. The following brief phyletic hypothesis is offered here as one possible interpretation of the facts presented: The 16-paired species are amphidiploid hybrids of 8-paired species. Although no 8-paired *Nicotiana* species are known today, postulation of their existence in past time is not unreasonable; species of 9 and 10 pairs, respectively, constitute the ALATAE, a related section of *Nicotiana* and in *Petunia*, a genus which borders closely upon the same section, there are 7- and 9-paired species. The 24-paired Australian-South Pacific species are derived from hybridization of 16-paired species with 8-paired (not necessarily the same 8-paired species in each case), followed again by amphidiploidy. Either a still higher level of hybridization, namely, 24-paired species with 16-paired, followed by selfing or backcrossing, but more readily by the latter, is responsible for initiation of a series of numbers between these two extremes, or possibly in the course of hybridization of 16-paired with 8-paired species sesquidiploidy (cf. Webber[18]) has thus functioned.

It is intended that papers published elsewhere will treat in greater detail of the external morphology, geographical distribution, chromosome morphology, and hybrid behavior in the SUAVEOLENTES. Hypotheses bearing upon the phylesis of these species will be more critically examined at that time.

Summary.—The fifteen Australian-South Pacific species of *Nicotiana*, or the SUAVEOLENTES section of the subgenus PETUNIOIDES, are a geographically isolated group of complexly interrelated herbaceous species, the specific differences of which generally are of small magnitude. Thirteen have been studied cytologically. Four 16-paired species, one 18-, two 19-, two 20-, one 21-, one 22- and two 24-paired occur. Usually a number of chromosomes with subterminal centromere increases at the expense of those with median or submedian, as chromosome number increases. Twenty-six hybrids involving twelve species show that considerable to high pairing between species obtains throughout in first meiotic metaphase. This indicates a common origin. A phyletic cytological hypothesis

involving a base number of 8 and repeated hybridization in part accompanied by chromosome doubling is offered as a possible interpretation of chromosome number, morphology and behavior in the SUAVEOLENTES.

* To Dr. T. H. Goodspeed, Professor of Botany and Director of the Botanical Garden, University of California, the author expresses her sincere gratitude both for counsel and for extending every research facility, including seeds of most of the *Nicotiana* cultures investigated. In addition she acknowledges with pleasure seeds sent her by E. E. Cheel, J. M. Black, Professor J. B. Cleland and Hazel Pearce.

† Kostoff[10,11] has given a 32-paired race which he has grown the designation "*N. Eastii*." The author does not consider this and another 32-paired race sufficiently distinct from *N. suaveolens* morphologically to warrant separation.

[1] Goodspeed, T. H. (in press), *Univ. Calif. Publ. Bot.*, **18**, 335*ff* (1945*a*).

[2] Goodspeed, T. H., *Proc. Nat. Acad. Sci.*, **19**, 649–653 (1933).

[3] Wheeler, H-M., *Univ. Calif. Publ. Bot.*, **18**, 45–68 (1935).

[4] Goodspeed, T. H. (in press), *Univ. Calif. Publ. Bot.*, **18** (1945*b*).

[5] Goodspeed, T. H., and Clausen, R. E., *Univ. Calif. Publ. Bot.*, **11**, 117–125 (1927).

[6] East, E. M., *Bibliogr. Gen.*, **4**, 243–320 (1928).

[7] Goodspeed, T. H., *Univ. Calif. Publ. Bot.*, **17**, 369–398 (1934).

[8] Kostoff, D., Dogadkina, N., and Tichonova, A., *Compt. Rend. (Doklady) Acad. Sci. U.S.S.R.*, **3**, 401–404 (1935).

[9] Kostoff, D., *Cytogenetics of the Genus Nicotiana*, Russian and English, Sofia (1941–1943), 1071 pp.

[10] Kostoff, D., *Curr. Sci.*, **8**, 110–112 (1939).

[11] Kostoff, D., *Compt. Rend. (Doklady) Acad. Sci. U.S.S.R.*, **22**, 267–269 (1939).

[12] Wheeler, H-M., *A Cyto-genetic Survey of the Species of Nicotiana Indigenous to Australasia*, Ph.D. thesis on file, Univ. Calif. Library (1934).

[13] Lewitsky, G. A., *Trudy Prikl. Bot., Gen. i Selek.*, **27**, 187–240, Russian and English (1931).

[14] Sarana, M. O., *Krasnodar, Vses. n.-i. Inst. Tab. Prom., Trudy*, No. **110**, 179–190, Russian with English summary (1934).

[15] Goodspeed, T. H. (in press), *Bot. Rev.*, **11** (1945).

[16] Hollingshead, L., *Univ. Calif. Publ. Bot.*, **11**, 257–264 (1929).

[17] Kostoff, D., *Genetica*, **22**, 215–230 (1940).

[18] Webber, J. M., *Univ. Calif. Publ. Bot.*, **11**, 319–354 (1930).

ON A CLASS OF HERMITIAN TRANSFORMATIONS CONTAINING SELF-ADJOINT DIFFERENTIAL OPERATORS

BY HANS LUDWIG HAMBURGER

UNIVERSITY COLLEGE, SOUTHAMPTON, ENGLAND

Communicated April 10, 1945

1. We consider the space $\mathfrak{L}_2(\Delta)$, that is, the space of all real and complex functions $f(x)$ defined in a certain (finite or infinite) closed interval Δ of the x-axis and such that $|f(x)|^2$ is Lebesque-integrable in Δ. The scalar product is then given by the relation

$$(f, g) = \int_\Delta f(x)\overline{g(x)}\, dx.$$

We denote by s any interior point of Δ, and introduce a family of projectors P_s in $\mathfrak{L}_2(\Delta)$ by writing

$$P_s f(x) = \begin{cases} f(x) & \text{for } x \leq s \\ 0 & \text{for } x > s \end{cases} \quad (x \in \Delta).$$

Now let H be a closed Hermitian transformation with domain $\mathfrak{D}$ contained in $\mathfrak{L}_2(\Delta)$, and denote by $\mathfrak{D}_s(\mathfrak{D}_s \subset \mathfrak{D})$ the set of all $f(x)$, such that both $f(x)$ and $P_s f(x)$ are elements of $\mathfrak{D}$. Clearly $\mathfrak{D}_s$ as well as $\mathfrak{D}$ is a linear manifold in $\mathfrak{L}_2(\Delta)$. If, moreover, $\mathfrak{D}_s$ is everywhere dense in $\mathfrak{D}$, then there exists a contraction ${}^sH$ of H with domain $\mathfrak{D}_s$. We further write as usual H^* for the adjoint of H, $H_\lambda = H - \lambda 1$, and $H^*_\lambda = H^* - \lambda 1$.

2. In this Note we are concerned with the class of Hermitian transformations H in $\mathfrak{L}_2(\Delta)$ satisfying the following conditions:

(α) *H is a closed Hermitian transformation*[1] *of deficiency index (m, m) with domain $\mathfrak{D}$.*

(β) *If $\{\varphi_\mu\}$ and $\{\psi_\mu\}$ are two sets of m linearly independent eigen solutions of H^* corresponding to the eigen values i and $-i$, respectively, so that $H^*_i\varphi_\mu = 0$, $H^*_{-i}\psi_\mu = 0$, $(\mu = 1, 2, \ldots, m)$, then the two sets of m elements $\{(P_t - P_s)\varphi_\mu\}$ and $\{(P_t - P_s)\psi_\mu\}$ are both linearly independent for every pair of numbers $s \neq t$ of Δ.*

(γ) *For every s interior to Δ, $\mathfrak{D}_s$ is everywhere dense in $\mathfrak{D}$; the contraction ${}^sH$ of H is of deficiency-index $(2m, 2m)$.*

(δ) *P_s and ${}^sH$ are commutative, so that for every f of $\mathfrak{D}_s$*

$${}^sHP_sf = P_s{}^sHf.$$

Such a transformation will be called an m^{th} order transformation of class P.

3. MAIN THEOREM. *Let H be an m^{th} order transformation of class P in $\mathfrak{L}_2(\Delta)$. If H satisfies the two additional conditions:*

(ϵ) *The differential quotients up to the m^{th} order of the m eigen solutions $\varphi_\mu = \varphi_\mu(x)$ defined in (β) exist at every point x of Δ,* and the $\psi_\mu = \psi_\mu\alpha$ are continuous in Δ.

(ζ) *$\mathfrak{D}_s \cdot \mathfrak{D}_t$ is everywhere dense in $\mathfrak{D}$ for every pair of numbers s, t contained in the interior of Δ.*

Then H^* can be represented by a self-adjoint differential operator[2] *of m^{th} order.*

Remark.—It is nearly obvious that the converse of the Theorem is true, i.e., that if H^* is an m^{th} order self-adjoint differential operator defined, in Δ, then its adjoint H satisfies the conditions (α) to (ζ).

The proof of this theorem will be given in a later paper. It is based entirely on the methods developed in [1] and [2], Chapter I.

4. There exist m^{th} order transformations of class P other than the self-

adjoint differential operators, which, naturally, cannot satisfy both of the conditions (ϵ) and (ζ). The later paper will give a method of obtaining all transformations of class P, and of representing them by means of certain integral equations.

We consider in particular the subclass $\mathfrak{S}_m$ of all those m^{th} order transformations H of class P whose m eigen solutions $\varphi_\mu(x)$, as defined in (β), have differential quotients up to the order $2m$ at every point x of Δ. (Hence the transformations of $\mathfrak{S}_m$ satisfy (α) to (ϵ) but not (ζ).) Let $\sigma(x)$ be any real, m-times differentiable function, $\sigma \equiv 0$ included, let $L_m(f)$ and $D_n(f)$ denote linear differential operators of order m and n, respectively ($n \leq m - 1$), and L_m^* and D_n^* Lagrange's adjoint operators. Then we can state the following:

THEOREM. *The subclass $\mathfrak{S}_m$ coincides with the class of all H, such that for all m-times differentiable functions $g(x)$ the relation $H_\lambda^* f = g$ can be represented in the form*[3]

$$L_m((\cos\sigma - \lambda\sin\sigma)f) - D_n((\sin\sigma + \lambda\cos\sigma)f) = L_m(g\sin\sigma) + D_n(g\cos\sigma) \tag{1}$$

subject to the following three conditions for L_m, D_n and σ:

(i) *The linear operator $L_m D_n^*$ of order $m + n$ is self-adjoint.*

(ii) *The differential equation*

$$L_m((\cos\sigma - i\sin\sigma)y) - D_n((\sin\sigma + i\cos\sigma)y) = 0$$

has a fundamental system of m solutions $\varphi_\mu(x)$ which are all elements of $\mathfrak{L}_2(\Delta)$.

(iii) *The differential equation*

$$L_m(z\sin\sigma) + D_n(z\cos\sigma) = 0$$

has singularities at both end-points of Δ, such that there exists no particular integral z of it for which $|z|^2$ is Lebesque-integrable in any sub-interval of Δ containing at least one of the end-points of Δ.

It can readily be shown that transformations of the subclass $\mathfrak{S}_m$ exist for all m, with $\sigma(x) \equiv 0$ as well as with $\sigma(x) \not\equiv 0$ in Δ. If $m \geqslant 2$, we can choose $n \not\equiv 0$.

5. If we wish to obtain all real transformations of the subclass $\mathfrak{S}_m$ defined in **4** we only have to take those transformations defined by (1) where the operators L_m and D_n are both real. Then owing to condition (i) of the above Theorem, $m - n$ is even. Hence for $m = 2$ the transformations defined by

$$L_2((\cos\sigma - \lambda\sin\sigma)f) - (\sin\sigma + \lambda\cos\sigma)f = L_2(g\sin\sigma) + g\cos\sigma,$$

L_2 being a self-adjoint differential operator, are the only real transformations of the subclass $\mathfrak{S}_2$. If $\sigma \equiv 0$, we obtain the case of the Main-Theorem, $L_2(f) - \lambda f = g$.

6. In order to include the singular case dealt with first in a series of papers by H. Weyl[4] where the number of linearly independent solutions of $L_m(y) = \lambda y$, (λ non-real), belonging to $\mathfrak{L}_2(\Delta)$ is less than the order of $L_m(y)$, we introduce the following generalization of the transformations of class P:

Definition.—Let Δ' be any interval $a' \leqslant x \leqslant b'$ whose end-points a' and b' are interior points of Δ, and write $P_\Delta' = P_b' - P_a'$. Then H, with domain $\mathfrak{D}$ in $\mathfrak{L}_2(\Delta)$, is an m^{th} order transformation of the generalized class P, if $P_\Delta'HP_\Delta'$ with domain $\mathfrak{D}\cdot\mathfrak{L}_2(\Delta')$ is an m^{th} order transformation[5] of class P in $\mathfrak{L}_2(\Delta')$ for every Δ'. This definition implies, firstly, that $P_\Delta'HP_\Delta'$ satisfies condition (ζ) for every Δ' and every pair of points s, t contained in Δ', and, secondly, that there exists a set of m linearly independent functions $\varphi_\mu(x)$ defined in every interior point x of Δ such that for every Δ'

$$P_\Delta'\varphi_\mu \in \mathfrak{L}_2(\Delta'), \quad (P_\Delta'H_iP_\Delta')^*\varphi_\mu = 0 \quad (\mu = 1, 2, \ldots, m). \qquad (2)$$

Hence the Main Theorem of **3** in the generalized case takes the following form:

THEOREM. *If H is an m^{th} order transformation of the generalized class P in $\mathfrak{L}_2(\Delta)$, and if the functions $\varphi_\mu(x)$ defined in (2) satisfy condition (ϵ), then H^* can be represented by a self-adjoint differential operator of m^{th} order.*

7. Finally, we define a class P of Hermitian transformations H given in any abstract Hilbert space $\mathfrak{H}$. Let $P(s)$ be a resolution of the identity defined in $\mathfrak{H}$ for every point s of a (finite or infinite) interval Δ, such that

$$1 = \int_\Delta dP_{(s)}$$

Let the Hermitian operator K defined in $\mathfrak{H}$ by the relation

$$K = \int_\Delta s dP(s)$$

be such that (i) K has no eigen value, (ii) K has simple spectrum, and (iii) no point of Δ belongs to the resolvent set of K. Then a closed Hermitian transformation H in $\mathfrak{H}$ is an m^{th} order transformation of class P, if there exists any resolution of the identity $P(s)$ in $\mathfrak{H}$ with the properties (i), (ii) and (iii), with respect to which H satisfies conditions (α) to (δ).

In order to obtain a theorem for the class P in $\mathfrak{H}$ which corresponds to the Main Theorem for the class P in $\mathfrak{L}_2(\Delta)$, we consider the one-one mapping of $\mathfrak{H}$ on to $\mathfrak{L}_2(\Delta)$ carried out[6] by means of a suitable element g_0 of $\mathfrak{H}$ such that any element $f(x)$ of $\mathfrak{L}_2(\Delta)$ and the corresponding f of $\mathfrak{H}$ are connected by the relation

$$f = \int_\Delta f(s) d(P(s)g_0). \qquad (3)$$

Then the Main Theorem takes the following form:

THEOREM. *Let H be an m^{th} order transformation of class P in $\mathfrak{H}$ which*

also satisfies condition (ζ), *and let* $\{\varphi_\mu\}$ *and* $\{\psi_\mu\}$ *be each a set of m linearly independent eigen solutions of* H^* *corresponding to the eigen values i and* $-i$, *respectively. If we can find a mapping of* $\mathfrak{H}$ *on to* $\mathfrak{L}_2(\Delta)$ *of the form* (3) *which maps the sets of elements* $\{\varphi_\mu\}$ *and* $\{\psi_\mu\}$ *on to sets of functions* $\{\varphi_\mu(x)\}$ *and* $\{\psi_\mu^{(x)}\}$ *satisfying condition* (ϵ), *then this mapping takes H into a self-adjoint differential operator of* m^{th} *order defined in* $\mathfrak{L}_2(\Delta)$.

In the same way the generalized class P of m^{th} order in $\mathfrak{L}_2(\Delta)$ can be extended to a generalized class P of m^{th} order in $\mathfrak{H}$. We omit the details.

[1] Hamburger, H. L., *Quart. Jour. Math.* (Oxford Series), **13**, 117–128 (1942).
[2] Hamburger, H. L., *Ann. Math.*, **45**, 59–99 (1944).
[3] Stone, M. H., *Linear transformations in Hilbert space*, New York, 1932.
[4] Weyl, H., *Math. Annalen*, **68**, 220–269 (1910).
[5] Weyl, H., *Göttinger Nachrichten*, pp. 442–467, 1910.

[1] See the definition of deficiency-index in [3] p. 81, Definition 2.21; p. 338, Definition 9.1.

[2] Here we use the word self-adjoint with the meaning which is implicit in Lagrange's general theory of ordinary linear differential equations. This differs essentially from its use in von Neumann's Theory (see [3], p. 50, Definition 2.11; p. 347, the last four lines), where a self-adjoint operator denotes a closed Hermitian operator of deficiency-index (0, 0).

[3] The equation (1) determines the relation $H_\lambda^* f = g$ only in a subdomain $\hat{\mathfrak{D}}$ of $\mathfrak{D}^*$, the domain of H^*. $\hat{\mathfrak{D}}$ is everywhere dense in $\mathfrak{D}^*$. If $\hat{H}_\lambda$ denotes the contraction of H_λ^* defined in $\hat{\mathfrak{D}}$, then it can be shown that $\hat{H}_\lambda$ is not closed, and that H_λ^* is the closure of $\hat{H}_\lambda$.

[4] See [4] and [5]; see also [3], pp. 458–498.

[5] If H, as a transformation in $\mathfrak{L}_2(\Delta)$, is itself of deficiency-index (m, m), then we are reduced to the case dealt with in **2** and **3**. In order to obtain something new, we have to assume that H in $\mathfrak{L}_2(\Delta)$ is of deficiency-index (m', m'), where $0 \leq m' < m$.

[6] See [3], p. 226, Theorem 6.2; pp. 275–277, Theorems 7.9 and 7.10.

NEW TYPES OF RELATIONS IN FINITE FIELD THEORY (SECOND PAPER)

BY H. S. VANDIVER

DEPARTMENT OF PURE MATHEMATICS, UNIVERSITY OF TEXAS

Communicated June 4, 1945

In the first paper published in these PROCEEDINGS under the above title,[1] the writer obtained some new types of relations in finite field theory. Here we shall proceed further along these general lines.

Mitchell[2] considered the relation

$$ax^m = by^m + 1 \tag{1}$$

in a finite field of order p^{2n}, p prime, including the given coefficients a and

b, with $p^{2n} - 1 = mc$, and there exists an n_1 such that $p^{n_1} \equiv -1 \pmod{m}$ and $n = sn_1$. He found that if a and b were each mth powers in $F(p^{2n})$, then the number of solutions (x^m, y^m) of (1) in this case is

$$\frac{1}{m^2}\left(p^{2n} + (m-1)(m-2)(-1)^{s-1}p^n - 3m + 1\right). \tag{2}$$

Also if a is an mth power and b is not, or conversely, then the number of solutions (x^m, y^m) of (1) is

$$\frac{1}{m^2}\left(p^{2n} - (m-2)(-1)^{s-1}p^n - m + 1\right). \tag{3}$$

If neither a nor b nor a/b is an mth power in the field, he found the number of solutions to be

$$\frac{1}{m^2}\left(p^{2n} + 2(-1)^{s-1}p^n + 1\right). \tag{4}$$

We now modify the method employed by the writer in another paper[3] to obtain other relations involving the number of solutions of (1), and we find as in the proof of Theorem I of that paper that in the field $F(p^{2n})$ we have, if N is the number of solutions of (1), and if

$$h = \frac{t(p^n - 1)}{c}, \tag{4a}$$

then

$$N = -c^2 \sum_{k=1}^{h} \sum_{r=0}^{k} \binom{kc}{rc} a^{mc-kc} b^{rc}$$

in the field. Comparing this with (2), (3), and (4), we have, noting that $p = 0$ in $F(p^{2n})$, the

THEOREM I. *In a finite field of order p^{2n}, with $p^{2n} - 1 = mc$, and such that there exists an n_1 such that $p^{n_1} \equiv -1 \pmod{m}$ then, if h is defined as in (4a), the expression*

$$\sum_{k=1}^{h} \sum_{r=0}^{k} \binom{kc}{rc} a^{mc-kc} b^{rc}$$

equals $(3m-1)$, $(m-1)$ or -1 in $F(p^{2n})$ according as (1) a and b are each m*th powers; (2) a is an* m*th power and b is not or conversely; (3) neither a nor b are* m*th powers, and also $a^c/b^c \neq 1$.*

In another paper,[4] the writer obtained the results that if p belongs to the exponent n, modulo l, and $p^n = 1 + cl$, where p and l are odd primes and n is even, and if by definition

$$\binom{m}{n} = 0 \tag{4b}$$

where $n > m$, then

$$\sum_{t=0}^{\infty} \binom{c}{a + tl} \equiv 0 \pmod{p}$$

except when $2a \equiv c \pmod{l}$, in which case the right-hand member is unity. We shall here extend this relation by generalizing the argument employed for the original proof.

If $\zeta = e^{2i\pi/l}$ and $\mathfrak{p}$ is a prime ideal divisor of (p) in $k(\zeta)$, then, if $k \not\equiv 0 \pmod{l}$,

$$\left(\frac{1 + \zeta^k}{\mathfrak{p}}\right) = \left(\frac{\zeta^{k/2}}{\mathfrak{p}}\right)\left(\frac{\zeta^{k/2} + \zeta^{-k/2}}{\mathfrak{p}}\right) \tag{5}$$

where these symbols represent power characters in $k(\zeta)$. Since

$$\left(\frac{\zeta^{k/2} + \zeta^{-k/2}}{\mathfrak{p}}\right) = 1,$$

(5) gives (6) and (7), for s an integer ≥ 0.

$$(1 + \zeta^k)^{cs} \equiv \zeta^{cks/2} \pmod{\mathfrak{p}}. \tag{6}$$

$$(1 + \zeta^k)^{cs+r} \equiv \zeta^{cks/2}(1 + \zeta^k)^r \pmod{\mathfrak{p}},\ 0 \leq r < c. \tag{7}$$

Since (7) holds for each value of $k \not\equiv 0 \pmod{l}$, we may expand and, after reducing each ζ^k by the relation $(\zeta^k)^l \equiv 1$, collect terms of the left- and right-hand members separately, calling the sum of the coefficients of $(\zeta^k)^i$ on the left A_i, and the sum of the coefficients of $(\zeta^k)^i$ on the right B_i. Transposition will give the following array, where each dth line is the special case for $k = d$, and each coefficient $(A_i - B_i)$ is abbreviated to K_i: $i = 0, 1, \ldots, l - 1$.

$$\left.\begin{array}{l} K_0 + K_1 \quad + K_2 \quad + \ldots + K_{l-1} \equiv 0 \\ K_0 + K_1(\zeta) + K_2(\zeta)^2 + \ldots + K_{l-1}(\zeta)^{l-1} \equiv 0 \\ K_0 + K_1(\zeta^2) + K_2(\zeta^2)^2 + \ldots + K_{l-1}(\zeta^2)^{l-1} \equiv 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ K_0 + K_1(\zeta^{l-1}) + K_2(\zeta^{l-1})^2 + \ldots + K_{l-1}(\zeta^{l-1})^{l-1} \equiv 0 \end{array}\right\} \tag{8}$$

modulo $\mathfrak{p}$. Consider the determinant formed by the coefficients K_i. It is an alternant which may be expressed, aside from powers of ζ, as the product of binomials of the type ζ^{q-1}, when $0 < q < l$. Any ideal $(\zeta^q - 1)$ is a prime ideal divisor of (l), hence is prime to (p). Therefore, $K_i = A_i - B_i \equiv 0 \pmod{p}$; $i = 0, 1, \ldots, l - 1$. Substitution of the actual values of A_i and B_i, obtained by the binomial expansion, gives a formula which may be expressed in the form (9), provided we define

$$\binom{0}{0} = 1. \tag{8a}$$

$$\sum_{t=0}^{\infty}\binom{cs+r}{a+tl} \equiv \sum_{t=0}^{\infty}\binom{r}{u+tl} \pmod{p}, \tag{9}$$

where a and u are integers, such that $u \equiv a - \frac{cs}{2} \pmod{l}$, $0 \leqq u \leqq l$, and $0 \leqq a < l$.

Also when, as before, $k \not\equiv 0 \pmod{l}$, then

$$\left(\frac{1-\zeta^k}{\mathfrak{p}}\right) = \left(\frac{\zeta^{k/2}}{\mathfrak{p}}\right)\left(\frac{\zeta^{k/2}-\zeta^{-k/2}}{\mathfrak{p}}\right). \tag{10}$$

Since

$$\left(\frac{\zeta^{k/2}-\zeta^{-k/2}}{\mathfrak{p}}\right) = 1,$$

(10) gives

$$(1-\zeta^k)^{cs} \equiv \zeta^{cks/2} \pmod{\mathfrak{p}}. \tag{11}$$

Hence

$$(1-\zeta^k)^{cs+r} \equiv \zeta^{cks/2}(1-\zeta^k)^r \pmod{\mathfrak{p}},\ 0 \leqq r < c. \tag{12}$$

Whence, as before, except that here $r > 0$,

$$\sum_{t=0}^{\infty}\left[(-1)^{a+tl}\binom{cs+r}{a+tl}\right] \equiv \sum_{t=0}^{\infty}\left[(-1)^{u+tl}\binom{r}{u+tl}\right] \pmod{p}, \tag{13}$$

where a and u are integers such that $u \equiv a - \frac{cs}{2} \pmod{l}$, $0 \leqq u < l$, and $0 \leqq a < l$. Where, however, $r = 0$, then since (11) holds for each value of $k \not\equiv 0 \pmod{l}$, we may expand the left-hand member and set up the following array, where each dth line is the special case for $k = d$, and where each A_i is the sum of the coefficients of $(\zeta^k)^i$, each ζ^k being reduced by the relation $(\zeta^k)^l \equiv 1$:

$$\left.\begin{array}{l}
A_0 + A_1 + A_2 + A_3 + \ldots + A_{l-1} \equiv 0,\\
A_0 + A_1\zeta + A_2\zeta^2 + A_3\zeta^3 + \ldots + A_{l-1}\zeta^{l-1} \equiv \zeta^{cs/2},\\
A_0 + A_1(\zeta^2) + A_2(\zeta^2)^2 + A_3(\zeta^2)^3 + \ldots + A_{l-1}(\zeta^2)^{l-1} \equiv (\zeta^2)^{cs/2},\\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
A_0 + A_1(\zeta^{l-1}) + A_2(\zeta^{l-1})^2 + A_3(\zeta^{l-1})^3 + \ldots + A_{l-1}(\zeta^{l-1})^{l-1} \equiv\\
\qquad (\zeta^{l-1})^{cs/2},
\end{array}\right\} \tag{14}$$

modulo $\mathfrak{p}$. Adding these congruences gives

$$lA_0 \equiv -1 \pmod p, \text{ if } \frac{cs}{2} \not\equiv 0 \pmod l.$$

To determine, in general, the value of any A_i, multiply each dth row of the original array by $(\zeta^{d-1})^{-i}$. The addition of the congruences of the resulting array gives

$$lA_i \equiv -1 \pmod p, \text{ if } \frac{cs}{2} - i \not\equiv 0 \pmod l; \tag{15}$$

$$lA_i \equiv l - 1 \pmod p, \text{ if } \frac{cs}{2} - i \equiv 0 \pmod l. \tag{16}$$

The relations (9), (13), (15), and (16) give

THEOREM II. *If p belongs to the exponent n, modulo l, and $p^n = 1 + cl$, where p and l are odd primes and n is even, and if s is any integer > 0 and a and u are integers, such that*

$$u \equiv a - \frac{cs}{2} \pmod l,\ 0 \leqq u < l, \text{ and } 0 \leqq a < l,$$

then

$$\sum_{t=0}^{\infty} \binom{cs + r}{a + tl} \equiv \sum_{t=0}^{\infty} \binom{r}{u + tl} \pmod p;$$

for r an integer $\geqq 0$;

$$\sum_{t=0}^{\infty} \left[(-1)^{a+tl} \binom{cs + r}{a + tl} \right] \equiv \sum_{t=0}^{\infty} \left[(-1)^{u+tl} \binom{r}{u + tl} \right] \pmod p,$$

if r is an integer > 0;

$$l \left\{ \sum_{t=0}^{\infty} \left[(-1)^{a+tl} \binom{cs}{a + tl} \right] \right\} \equiv -1 \pmod p,$$

if $cs/2 - a \not\equiv 0 \pmod l$
and

$$l \left\{ \sum_{t=0}^{\infty} \left[(-1)^{a+tl} \binom{cs}{a + tl} \right] \right\} \equiv l - 1 \pmod p,$$

if $cs/2 - a \equiv 0 \pmod l$.

As an application of Theorem II, we shall now prove the following:

THEOREM III. *If F is a finite field containing b and of order $p^{2n} = 1 + cl$ (where p and l are odd primes) and if p belongs to the exponent $2n$ (modulo l) and $bxy \neq 0$, and where N denotes the number of solutions (x^c, y^l) in F of the equation*

$$x^c + by^l + 1 = 0, \tag{17}$$

then

$$N \equiv 1 \pmod{p}, \ \textit{if } (c, l) = 1; \tag{18}$$

$$N \equiv 0 \pmod{p}, \ \textit{if } (c, l) = l \textit{ and } b^c \not\equiv 1 \pmod{p}; \tag{19}$$

$$N \equiv l \pmod{p}, \ \textit{if } (c, l) = l \textit{ and } b^c \equiv 1 \pmod{p}. \tag{20}$$

By Theorem II of another paper,[5] for the equation (17)

$$N = -hi \sum_{k=1}^{l} \sum_{\rho=0}^{\gamma} \binom{ki}{h\rho} b^{p^{2n}-1-ki}(-1)^{ki},$$

where $h = l$, $i = c$, and $\gamma = [ki/h]$.

$$N = -lc \sum_{k=1}^{l} \sum_{\rho=0}^{\gamma} \binom{kc}{l\rho} b^{-kc}(-1)^{kc}. \tag{21}$$

Now $p^{2n} - 1$ is even and l is odd, so c is even. By application of Theorem II of this paper,

$$N \equiv \sum_{k=1}^{l} \sum_{\rho=0}^{\infty} \binom{0}{u + \rho l} b^{-kc} \pmod{p}, \tag{22}$$

where u is an integer, such that

$$0 \leqq u < l \text{ and } u \equiv -\frac{ck}{2} \pmod{l}.$$

Suppose $(c, l) = 1$. Then $u \not\equiv 0 \pmod{l}$, except when $k = l$, in which case $u = 0$, and we find, using (4*b*) and (8*a*), $N \equiv 1 \pmod{p}$.

Suppose $(c, l) = l$. Then $u = 0$ for every value of k, and (22) reduces to

$$N \equiv \sum_{k=1}^{l} b^{-kc} \pmod{p},$$

from which we obtain (19) and (20). This proves Theorem III.

In (17), x^c cannot take on more than l distinct values, and also we cannot have two solutions, (x^c, y_1^l) and (x^c, y_2^l), with $y_1^l \neq y_2^l$. Likewise, y^l cannot take on more than c distinct values, and for any two solutions, (x_1^c, y^l) and (x_2^c, y^l), $x_1^c = x_2^c$. Hence, the total number of solutions can never exceed either c or l, and (20) becomes $N = l$. In (19), $N = 0$, if either $c < p$ or $l < p$, and in (18), $N = 1$, if either $c < p + 1$ or $l < p + 1$.

[1] Vandiver, H. S., These PROCEEDINGS, 31, 50–54 (1945).

[2] Mitchell, *Ann. Math.*, 18, 120 (1917).

[3] These PROCEEDINGS, 30, 362–367 (1944).

[4] *Ann. Math.*, (2), 28, 330–332 (1927).

[5] These PROCEEDINGS, 31, 170–175 (1945).

ON ALGEBRAIC LIE ALGEBRAS

By Claude Chevalley and Hsio-Fu Tuan

Department of Mathematics, Princeton University

Communicated June 15, 1945

A subgroup A of the full linear group $GL(n, C)$ in n variables (C being the field of complex numbers) is called *algebraic* if the condition for a matrix σ of $GL(n, C)$ to belong to A can be expressed by a system of algebraic equations in the coefficients of σ. Clearly A is then a complex Lie group, and it is of interest to determine the Lie algebras which can be considered as Lie algebras of algebraic Lie groups. The question has been solved in a certain sense by Maurer;[1] here we propose to resume the study from a different point of view.

One of us[2] has defined the notion of a *replica* Y of a matrix X of degree n as any matrix Y which admits as its invariants all the tensor-invariants of X (where, in defining tensor-invariants, X is meant to be the symbol of an infinitesimal, not a finite, transformation; thus the vector invariants are the vectors $\mathfrak{x}$ such that $X\mathfrak{x} = 0$, and not $X\mathfrak{x} = \mathfrak{x}$. Now, let K be a field of characteristic 0 and let $\mathfrak{gl}(n, K)$ be the Lie algebra of matrices of degree n with coefficients in K. A subalgebra $\mathfrak{g}$ of $\mathfrak{gl}(n, K)$ will be called *algebraic* if every replica of a matrix $X \epsilon \mathfrak{g}$ still belongs to $\mathfrak{g}$.

If K is the field C of complex numbers, it is easily seen that the Lie algebra spanned by all the replicas of a matrix X is the Lie algebra of the smallest algebraic subgroup of $GL(n, C)$ to contain the one-parameter group generated by X. It follows immediately that the Lie algebra of an algebraic group of matrices is algebraic. The converse is true and can be seen from the following considerations. Let $\mathfrak{g}$ be an algebraic Lie algebra; then we may consider the elements of $\mathfrak{g}$ as operating on a certain vector space $\mathfrak{M}$, and therefore also on the tensor spaces $\mathfrak{T}_{r,s}$ which can be constructed on $\mathfrak{M}$ ($\mathfrak{T}_{r,s}$ is the space of r times contravariant and s times covariant tensors). Now we consider all pairs of vector subspaces $(\mathfrak{P}, \mathfrak{Q})$ of all $\mathfrak{T}_{r,s}$ such that $\mathfrak{Q} \subset \mathfrak{P}$, $X_{r,s}(\mathfrak{P}) \subset \mathfrak{Q}$ for all $X \epsilon \mathfrak{g}$ (where $X_{r,s}$ is the operation on $\mathfrak{T}_{r,s}$ which corresponds to X). It can then be proved that if a matrix X' is such that $X'_{r,s}(\mathfrak{P}) \subset \mathfrak{Q}$, then X' belongs to $\mathfrak{g}$. It is clear that a condition of the form $X'_{r,s}(\mathfrak{P}) \subset \mathfrak{Q}$ can be expressed by algebraic equations on the coefficients of the matrices of the one-parameter group generated by X'. It then follows that *any algebraic Lie algebra over the field of complex numbers is the Lie algebra of an algebraic group of matrices.* Moreover, our method of proof gives an indication as to how to write down a system of finite equations which define the group.

Let $\mathfrak{g}$ be any subalgebra of $\mathfrak{gl}(n, K)$. Among all the algebraic Lie algebras to contain $\mathfrak{g}$, there clearly exists a smallest one, say $\mathfrak{g}^*$. It can be proved

that $\mathfrak{g}^*$ has the same derived algebra as $\mathfrak{g}$ itself and that every ideal in $\mathfrak{g}$ is also an ideal in $\mathfrak{g}^*$.

Let $\mathfrak{g}$ be any algebraic Lie algebra. Denote by $\mathfrak{h}$ the radical of $\mathfrak{g}$ (i.e., the largest solvable ideal in $\mathfrak{g}$) and by $\mathfrak{n}$ the largest ideal of $\mathfrak{g}$ composed only of nilpotent matrices. By Levi's theorem, $\mathfrak{g}$ is the direct sum of $\mathfrak{h}$ and of a semi-simple subalgebra $\mathfrak{f}$. It can be proved that $\mathfrak{h}$ is the direct sum of $\mathfrak{n}$ and of Abelian algebra $\mathfrak{a}$ whose matrices are semi-simple and commute with those of $\mathfrak{f}$.

Let $\mathfrak{g}$ be any subalgebra of $\mathfrak{gl}(n, K)$; then it can be shown that the derived algebra $\mathfrak{g}'$ is algebraic. Moreover, $\mathfrak{g}'$ can be "defined by its invariants," in the sense that any matrix which admits as its invariants all the common invariants of all matrices in $\mathfrak{g}'$ lies itself in $\mathfrak{g}'$. Our result applies in particular to any semi-simple Lie algebra $\mathfrak{g}$ of $\mathfrak{gl}(n, K)$, which is identical with its derived algebra $\mathfrak{g}'$. Moreover, our method of proof shows more generally that any subalgebra $\mathfrak{g}$ of $\mathfrak{gl}(n, K)$ whose radical is composed only of nilpotent matrices is algebraic and is defined by its invariants.

If A is any algebra (associative or not) over the field K, the derivations of A form a Lie algebra which is easily seen to be algebraic.

Finally, let it be mentioned that the notion of algebraic Lie algebras can be used with advantage in the exposition of the theory of semi-simple Lie algebras, notably in establishing Cartan's criterion of semi-simplicity and Lie's theorem on solvable Lie algebras. Barring the recourse to the algebraic closure of the basic field in the proof of Theorem 3 of the paper quoted above,[2] one obtains in this way a rational proof of Cartan's criterion and of Lie's theorem.

[1] Maurer, L., "Zur Theorie der continuerlichen, homogenen und linearen Gruppen," *Sitzungsber. d. Bayerischen Akad. Math. Phys. Classe*, **24**, 297–341 (1894).

[2] Chevalley, C., "A New Kind of Relationship between Matrices," *Amer. J. Math.*, **65**, 521–531 (1943).

TWO INTEGRAL EQUATIONS

BY H. BATEMAN

NORMAN BRIDGE LABORATORY OF PHYSICS, CALIFORNIA INSTITUTE OF TECHNOLOGY

Communicated May 19, 1945

Let $k(s, t) = g(s, t) \quad -c\, g(s, a - t)$ where $g(s, t)$ is a real continuous kernel for $0 \leq s, t \leq a$ and c is an arbitrary constant. The equations to be considered are

$$f(s) = \int_0^a k(s, t)\ F(t)dt \tag{1}$$

$$f(s) = F(s) - x\int_0^a k(s, t)F(t)dt \tag{2}$$

and the required function $F(t)$ is to be continuous when $f(s)$ is continuous. When this condition is supposed to be satisfied equation (1) may be written in the equivalent form

$$f(s) = \int_0^a g(s, t)[F(t) - c\,F(a - t)]dt$$

and when a continuous function $h(t)$ exists for which

$$f(s) = \int_0^a g(s, t)h(t)dt$$

a solution of the integral equation may be obtained by solving the functional equation

$$h(t) = F(t) - c\,F(a - t) \qquad 0 \leq t \leq a.$$

Since $h(a - t) = F(a - t) - c\,F(t)$ the solution should be

$$(1 - c^2)F(t) = h(t) + c.h(a - t).$$

If $c^2 = 1$ this last equation fails to determine $F(t)$ and this is just the case when the homogeneous integral equations

$$0 = \int_0^a k(s, t)E(t)dt, \qquad 0 = \int_0^a H(s)k(s, t)ds$$

have innumerable solutions. Thus, when $c = 1$ any solution of the functional equation $E(t) = E(a - t)$ which gives finite integrals is a solution of the first equation and if $L(t) = \int_0^a H(s)k(s, t)ds$, the second equation is satisfied whenever $L(t) = L(a - t)$.

For equation (2) the homogeneous integral equation

$$0 = E(s) - z\int_0^a k(s, t)E(t)dt$$

may be written in the form

$$E(s) = z\int_0^a g(s, t)[E(t) - c\,E(a - t)]dt \qquad 0 \leq s \leq a$$

Changing s into $a - s$ we have

$$c\,E(a - s) = cz\int_0^a g(a - s, t)[E(t) - c\,E(a - t)]dt$$

Hence the function $H(t) = E(t) - c\,E(a - t)$ is a solution of the quasi-adjoint[1] homogeneous integral equation

$$H(s) = z\int_0^a [g(s, t) - c\,g(a - s, t)]H(t)dt$$

When the equation for $E(t)$ is written in the form

$$E(s) - z\int_0^a g(s, t)E(t)dt + cz\int_0^a g(s, a - t)E(t)dt = 0$$

it is seen that

$$E(s) = -cz\int_0^a G(s, a - t)E(t)dt$$

where $G(s, t)$ is the solving kernel associated with $g(s, t)$. Now this is a

homogeneous integral equation for $E(t)$ and so $-cz$ is a root of the transcendental equation $\Delta(-cz, z) = 0$, where Δ is the determinantal function associated with the kernel $G(s, a - t)$ and is an entire function of $-cz$.

The kernel $k(s, t) = g(s, t) - c\,g(s, a - t)$ is only of a special type when $c^2 = 1$ for the two equations

$$k(s, t) = g(s, t) - c\,g(s, a - t),\ k(s, a - t) = g(s, a - t) - c\,g(s, t)$$

give the relation

$$k(s, t) + c\,k(s, a - t) = (1 - c^2)g(s, t)$$

from which the function $g(s, t)$ can be determined when c^2 is different from unity.

In the particular case in which

$$\begin{aligned} g(s, t) &= s(1 - t) \qquad s \leq t \qquad a = 1 \\ &= t(1 - s) \qquad t \leq s \end{aligned}$$

the integral equation (2) may be written in the form

$$f(s) = F(s) - x \int_0^1 g(s, t)[F(t) - c\,F(1 - t)]dt.$$

Differentiating twice with respect to s and using the known fact that $g(s, t)$ is the Green's function of the differential expression d^2y/ds^2 we obtain the two equations

$$\begin{aligned} f''(s) &= F''(s) + x[F(s) - c\,F(1 - s)] \\ f''(1 - s) &= F''(1 - s) + x[F(1 - s) - c\,F(s)]. \end{aligned}$$

Addition and subtraction gives the equations

$$f''(s) + f''(1 - s) = E''(s) + m^2E(s) \qquad m^2 = x(1 - c)$$

$$f''(s) - f''(1 - s) = B''(s) + n^2B(s) \qquad n^2 = x(1 + c)$$

$$E(s) = F(s) + F(1 - s), E(0) = E(1) = F(0) + F(1) = f(0) + f(1)$$

$$B(s) = F(s) - F(1 - s), B(0) = -B(1) = F(0) - F(1) = f(0) - f(1).$$

When $f(s) = 0$ we have simply

$$E''(s) + m^2E(s) = 0,\ E(0) = E(1) = 0$$

$$B''(s) + n^2B(s) = 0,\ B(0) = B(1) = 0.$$

Hence

$$E(s) = A \sin (ms) \text{ where } \sin (m) = 0,\ x(1 - c) = p^2\pi^2$$

$$B(s) = B \sin (ns) \text{ where } \sin (n) = 0,\ x(1 + c) = q^2\pi^2$$

where p and q are positive integers. The two equations

$$x(1-c) = p^2\pi^2, \; x(1+c) = q^2\pi^2$$

cannot generally be satisfied simultaneously although this can happen for special values of c. Usually when $x(1-c) = p^2\pi^2$, $B = 0$ and when $x(1+c) = q^2\pi^2$, $A = 0$.

In the general case in which $f(s)$ is not zero the solution of the integral equation (2) is

$$F(s) = f(s) + \frac{1}{2}\int_0^1 [U(s, t) + U(s, 1-t) + V(s, t) - V(s, 1-t)]f(t)dt$$
$$- \frac{1}{2}\int_0^s [m \sin\{m(s-t)\} + n\sin\{n(s-t)\}]f(t)dt$$
$$+ \frac{1}{2}\int_0^{1-s} [n \sin\{n(s-1+t)\} - m\sin\{m(s-1+t)\}]f(t)dt$$

where

$$U(s, t) = m \operatorname{cosec}(m) \sin(ms) \sin m(1-t)$$
$$V(s, t) = n \operatorname{cosec}(n) \sin(ns) \sin n(1-t).$$

Since

$$F(s) = f(s) + x\int_0^1 K(s, t)f(t)dt$$

where $K(s, t)$ is the solving kernel corresponding to $k(s, t)$, an expression for $K(s, t)$ may be written down.

The present example provides an interesting illustration of a method of solving a homogeneous integral equation of the first kind which was discussed about 36 years ago.[2]

When equation (2) with $f(s) = 0$ is written in the form

$$M(c)E(s) = \int_0^a [g(s, t) - c\,g(s, a-t)]E(t)dt$$

it seems likely that the values of c for which $M(c) = 0$ will be those for which the homogeneous integral equation

$$0 = \int_0^a [g(s, t) - c\,g(s, a-t)]E(t)dt$$

can be satisfied. Now in the present example the possible forms of $M(c)$ are

$$(1-c)/p^2\pi^2 \qquad \text{and } (1+c)/q^2\pi^2.$$

When $c = 1$ the function $\sin(ms)$ with an odd integral value of m/π satisfies the functional equation $E(t) = E(1-t)$ and when $c = -1$ the function $\sin(ns)$ with an even value of n satisfies the functional equation

$E(t) = -E(1 - t)$. The known solutions of the homogeneous integral equation of the first kind are thus actually furnished by this method if it is agreed that a relevant solution of the functional equation $E(t) = \pm E(1 - t)$ can be expressed as a linear combination of the functions of type $\sin (mt)$.

[1] When $g(s, t) = g(t, s)$ the equation for $H(s)$ is adjoint to the equation for $E(s)$.

[2] H. Bateman, *Trans. Camb. Phil. Soc.*, **21**, 195 (1909); *Messenger Math.*, **39**, 6–19, 182–191 (1909–1910). Integral equations containing a parameter have been discussed also by G. Barba, *Relat. Soc. Gideniae Catinensis Nat. Soc. Catania*, **66**, 3–10 (1934); N. Gioranescu, *Bull. Sci. Math.*, s. 2, **58**, 270–272 (1934); *Rend. Semin. Mat. Padova*, **5**, 81–98 (1934); C. Miranda, *Palermo Rend.*, **60**, 286–304 (1937); R. Iglisch, *Math. Ann.*, **117**, 129–139 (1939).

VITAL STATISTICS OF THE NATIONAL ACADEMY OF SCIENCES

By Edwin B. Wilson

Harvard School of Public Health

Communicated June 14, 1945

Under the title "Vital Statistics of the National Academy of Sciences," Raymond Pearl, about twenty years ago when acting as Chairman of a Committee appointed to consider matters concerned with the rules for election to the Academy, published a series of articles which still deserve attention.[1] He gave figures for the mean and median ages and the standard deviation of the age distribution for different periods in the history of the Academy. I reproduce his results with the addition of those for the 81 persons elected in the three years 1943–1945, and with columns showing the percentage under 40 years of age when elected and the number elected in the period.

Period	Mean Age	Median Age	Standard Deviation	Per Cent Under 40	Number Elected
Charter, 1863	51.7	51.3	11.1	16.7	48
1864–1883	44.5	41.3	10.2	45.5	95
1884–1904	46.5	45.1	8.5	20.0	65
1905–1924	50.5	49.5	8.1	8.4	213
1943–1945	52.1	51.8	7.6	3.7	81

It will be observed that the tendencies toward a higher mean age, a median nearer the mean, a smaller standard deviation about the mean, and a sharply decreasing percentage of really young persons elected, which Pearl so much deplored, have persisted and been intensified.

The average number of persons elected each year in the first twenty years after the charter was 4, in the next twenty years it was 3, in the twenty years 1905–1924 it was 11, and for the most recent three years it

was 27. At each time that there has been discussion of fixing the number who might well be elected each year, it has been held by some that the reason for the increasing average age at election was the increasing number of scientists of distinction who had not been elected to the Academy because of the small number who could be elected. There is no indication in the figures that the increased number elected has been sufficient to use up the accumulation of "older men who have not been elected"—if that is truly the reason for the increase in the mean age.

The first edition of *American Men of Science* in 1906 listed about 4000 whereas the seventh edition in 1944 listed 34,000. The number of members of the Academy in 1906 was 96 and in 1944 was 350. Whether the standard for inclusion in *American Men of Science* has declined or whether the fraction of really eminent scientists relative to the listings has declined would be difficult if not impossible to determine; but at any rate for each member of the Academy in 1906 there were about 42 persons in *American Men of Science* whereas for each Academician in 1944 there were about 97. The number of new starred names added to the volume in recent revisions has been at the average rate of 50 per annum.

Pearl computed a life-table for the membership of the Academy on the basis of the limited number (6273) of person-years of experience available from 1863 through 1924. A considerably larger experience is now available but for present purposes it will be satisfactory to use a standard life table.[2] The average expectation of life after election to the Academy as figured from this table and the age distribution of the 81 persons elected in 1943–1945 is 20.8 years whereas the expectation figured from the younger age distribution of 1864–1883 is 27.4, more than 30% higher. The ultimate size of the Academy if an average of 27 persons are elected each year, as during the past three years, on the basis of an average expectation of life of 20.8 years will be about 560.

It has long been known that annuitants have longer expectations of life than the insured. How closely to the 1937 Standard Annuity Table[3] the expectation of life of members of the Academy may prove to be when a sufficient experience has accumulated to give real stability to the results one cannot forecast, but it is interesting to note that from age 30 to age 70 the Annuity Table shows expectations of life about 2.4 years greater than those in the Standard Table Z based on insured lives, and would therefore give average expectations of life of 23.2 on the basis of the current age distribution at election, and of 29.8 on the basis of the age distribution of seventy years ago. Such increased expectations would mean proportionally increased numbers of members in the Academy if it elected a fixed number of persons annually until it reached a stationary population.

[1] These PROCEEDINGS, **11**, 752–768 (1925) and **12**, 258–261 (1926).

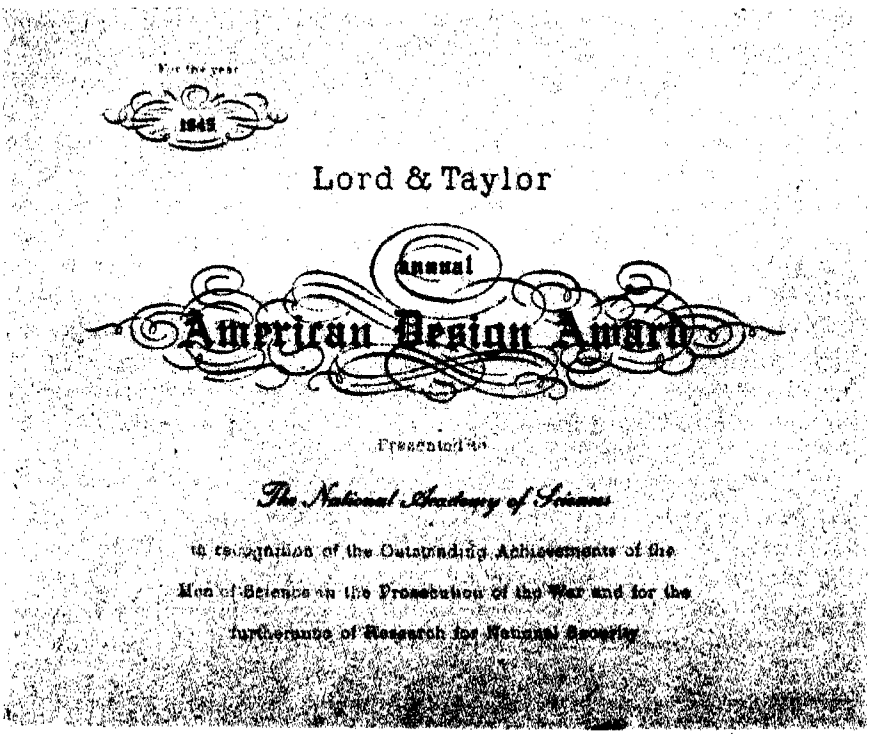
[2] After some discussion with Mr. Harold A. Grout, actuary of the John Hancock

Mutual Life Insurance Company, I have decided to use Table Z of the Mortality Tables and Life Insurance Statistics, which is an empirical table derived from the experience of insured lives, and which actually runs very close to the table Pearl computed for the Academy experience as of twenty years ago.

[3] The 1937 Standard Annuity Table, as well as the empirical Table Z, based on insured lives, was furnished to me by Mr. Grout, who informs me that the experience of the John Hancock Company with its male annuitants who purchased annuities from 1931 to 1943, inclusive, and who selected a non-refund type of annuity actually followed the table very closely.

AN AWARD TO THE ACADEMY

On April 19, 1945, the American Design Award for 1945, the diploma of which is herewith reproduced, was presented to the Academy and accepted by the President, Dr. F. B. Jewett, on behalf of the Academy, the



Research Council and all American science. The award carried a sum of \$25,000 available to the Academy without restriction. The donors expressed informally the hope, however, that the Academy would find ways to employ the money for purposes which in its judgment seemed likely to enhance national security.

Although the award comes to the Academy as the officially recognized leader in the organization of science in the service of the Government, special mention was made of six members of the Academy who have been directing important scientific activities in connection with war research, namely Messrs. Vannevar Bush, J. B. Conant, K. T. Compton, R. G. Harrison, J. C. Hunsaker and Dr. A. N. Richards.

THE VARIATION OF INFECTIVITY, II

BY EDWIN B. WILSON AND JANE WORCESTER

HARVARD SCHOOL OF PUBLIC HEALTH

Communicated June 13, 1945

We have discussed the variation of infectivity when certain curves are assumed as appropriate to describe the course of an epidemic.[1] One may also discuss the variation directly upon the data. Taking the infectivity as proportional to the reciprocal of the equilibrium value of m and using the law of mass action, we have

$$\text{infectivity} \propto \frac{1}{m} = \frac{1}{S}\left[\frac{C(t)}{C(t-\tau)}\right]^{1/p}. \tag{1}$$

The ratio of the case rates one incubation period apart, when that is one-half the period used in the tabulation ($\tau = {}^1/_2$), can be determined with adequate approximation[2] according to Soper as

$$\frac{C(t + {}^1/_4)}{C(t - {}^1/_4)} = \sqrt{\frac{K_0^1}{K_{-1}^0}}, \tag{2}$$

where the K's represent the numbers of cases during the month after and before $t = 0$.

Proceeding with (1) and (2) applied to Hedrich's data for measles in Baltimore for 32 years we have found for each month the value of $1/m$. Owing to the irregularities in the course of epidemics the values for $1/m$ for the Januaries (or any other specified month) are highly variable. The means of the group of 32 are, however, reasonably stable with standard errors of estimate of about 5%. In table 1 the values of the means of k/m have been entered after choosing k so that the average in any row is 1.00—the values thus indicating a relative seasonal index of infectivity. A comparison of lines 1 and 8 in the table shows that the Baltimore index has a much larger variation than the one Soper derived for Glasgow from a smaller series of years (as adjusted by interpolation to 12 months instead of 13 four-week periods per year). To see whether the index for Baltimore

would vary greatly for different samples taken from the 32 years, we have tabulated in lines 3, 4, 5, 6, respectively, the index as computed from the 15 epidemic years (September through the following August) with the largest numbers of cases, the 16 epidemic years with the smallest number of cases,[3] the first 16 years, and the last 16 years. The largest years and smallest years seem surely to have different seasonal distributions of the index, but there seems to be no notable difference between the seasonal variation of the index in the first and second halves of the 32-year period.

The degree of variation in the index will clearly depend on the value chosen for p. The indices so far mentioned were computed for $p = 1$. If p be assumed to have the value 2, we have the results tabulated in line 7 of table 1 which show a smaller variation than those in line 1. In line 9 are given for six months[4] values of an index computed for the six clear-cut epidemics for which values of p have been determined from the relation of total cases to peak cases, the variation of $1/m$ in each being computed with the appropriate value of p; the constancy of the results is to be expected.

There are striking differences in the seasonal distribution of the incidence of measles in different places and for different sets of years. This seasonal variation is shown, in table 2, by the percentage of cases falling

TABLE 1

VALUES OF THE MONTHLY MEANS OF THE INFECTIVITY STANDARDIZED TO HAVE A MEAN OF 1.00 FOR DIFFERENT PERIODS FOR BALTIMORE AND FOR $p = 1$ UNLESS OTHERWISE SPECIFIED

MONTH	SEPT.	OCT.	NOV.	DEC.	JAN.	FEB.	MAR.	APR.	MAY	JUNE	JULY	AUG.
Infect.[1]	0.79	1.10	1.46	1.25	1.26	1.17	1.12	1.05	0.95	0.73	0.60	0.53
(m)[2]	77	55	41	40	48	52	54	58	64	83	101	115
Infect.[3]	0.75	1.07	1.61	1.37	1.29	1.19	1.14	0.98	0.87	0.65	0.58	0.51
Infect.[4]	0.81	1.14	1.33	1.13	1.24	1.14	1.10	1.13	1.02	0.81	0.61	0.54
Infect.[5]	0.80	1.22	1.43	1.22	1.23	1.14	1.13	1.01	0.95	0.72	0.63	0.52
Infect.[6]	0.78	0.96	1.50	1.27	1.30	1.20	1.10	1.10	0.94	0.74	0.57	0.53
Infect.[7]	0.90	1.05	1.21	1.11	1.12	1.08	1.06	1.04	1.00	0.88	0.80	0.74
Infect.[8]	0.91	1.19	1.11	1.13	1.05	0.99	1.08	1.05	1.00	0.89	0.80	0.81
Infect.[9]	..	..	..	..	1.02	0.99	1.00	1.00	1.03	0.97	..	..

[1] Baltimore, 1900–1931. [2] Harmonic mean in thousands.
[3] Fifteen largest epidemic years. [4] Sixteen smallest epidemic years.
[5] Baltimore, 1900–1915. [6] Baltimore, 1916–1931.
[7] Baltimore, 1900–1931, $p = 2$. [8] Glasgow (Soper), 1905–1916.
[9] The six clear-cut Baltimore epidemics with p taken for each as determined.

TABLE 2

PERCENTAGE DISTRIBUTION OF MEASLES BY MONTHS

MONTH	SEPT.	OCT.	NOV.	DEC.	JAN.	FEB.	MAR.	APR.	MAY	JUNE	JULY	AUG.
Balt.[1]	0.4	0.7	2.3	5.1	10.3	14.9	19.2	20.0	16.5	7.4	2.4	0.7
Balt.[2]	0.3	0.6	2.5	5.7	10.7	14.8	19.4	20.0	16.4	6.6	1.9	0.5
Balt.[3]	1.1	1.3	2.6	3.3	5.3	9.5	15.5	18.9	20.3	14.0	6.0	2.2
Balt.[4]	0.5	1.0	3.3	7.7	11.6	14.7	18.7	17.8	13.9	7.1	2.8	0.8
Balt.[5]	0.3	0.4	1.6	2.9	9.3	15.0	19.6	21.8	18.7	7.7	2.2	0.6
Glas.[6]	2.5	4.8	8.2	12.3	14.4	12.4	10.9	9.6	10.1	7.4	4.8	2.7

[1] Baltimore, 1900–1931. [2] Fifteen largest epidemic years.
[3] Sixteen smallest epidemic years. [4] Baltimore, 1900–1915.
[5] Baltimore, 1916–1931. [6] Glasgow, 1901–1910.

within the month. It will be seen from lines 2 and 3 that the seasonal distribution of cases differs in the largest and smallest years as was the case with the index of infectivity; however, here there is also considerable difference in the seasonal distribution of measles during the first and second halves of the 32-year period whereas there seemed to be nothing really significant in the differences of the seasonal variation of the index of infectivity. It may be noted that the seasonal variation of measles in Glasgow is quite different from that of any of the Baltimore series in that there is a long flat peak in December to February, instead of a sharp peak in March to April or April to May, and a relatively high incidence throughout the summer.

There is some indication in the data that the seasonal variation of the index of infectivity may be in some way related to the manner of variation of the seasonal distribution of the disease—as that the seasonal variation of infectivity is greater when the seasonal variation in the incidence is greater—but we would not wish to make such generalizations short of a more extended comparison of results in a variety of places or short of more intensive study of the details of the calculation. For example, we have spoken, following Soper, of the seasonal variation of the infectivity k/m. This variation may be largely due to the rise and fall of the epidemic which cannot be expected to follow perfectly any assigned law such as (1) and the seasonal variation of infectivity might therefore be largely due to the seasonal distribution of the disease. Indeed we saw that for $p = 2$ it was possible to get an exact solution of the equation, namely,

$$S = m\ [\cosh \alpha - \sinh \alpha \tanh \alpha t/\tau],$$

with m strictly constant; but if now we apply Soper's formula assuming $p = 1$, we have

$$\text{infectivity} \propto \cosh \alpha - \sinh \alpha \tanh \alpha t/\tau$$

and hence, as thus calculated, the infectivity would diminish throughout the epidemic from an initial value of $\cosh \alpha + \sinh \alpha$ to a final value of $\cosh \alpha - \sinh \alpha$, although actually $1/m$ was constant by hypothesis.[5]

We have assumed p as fixed and m or $1/m$ to be determined from month to month by (1) and (2); we may examine the question of determining both m and p from the monthly data. Indeed, we have from (1) and (2)

$$\log m + \frac{1}{2p} \log (K_0^1/K_{-1}^0) = \log S. \qquad (3)$$

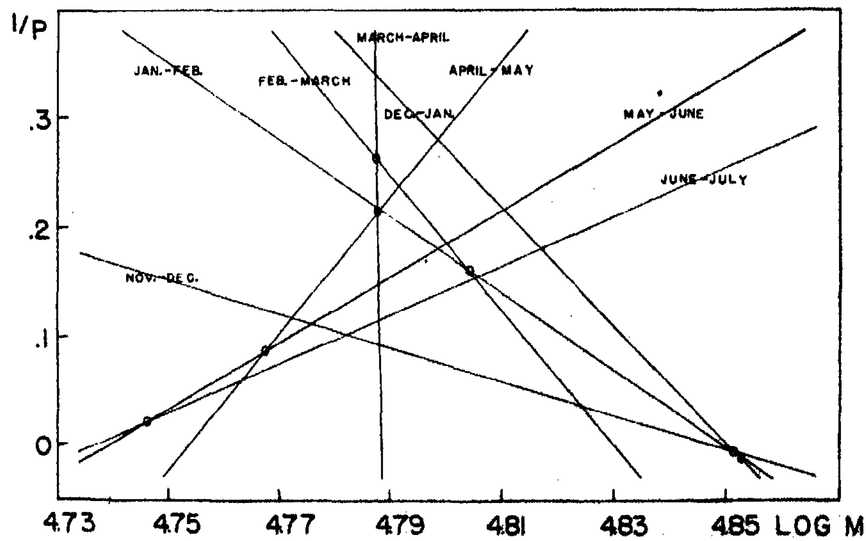
This imposes a linear condition upon $\log m$ and $1/p$ determinable from the cases in two successive months and the susceptibles a quarter of a month after the beginning of the second month. For the epidemic in Baltimore in the year 1912–1913, figure 1 gives the lines which result from the successive pairs of months. Owing to statistical fluctuations inevitable in

such phenomena as epidemics, the positions of the lines must be expected to show considerable instability. The intersections of successive pairs of lines determine the values of log m and $1/p$ that would be assigned in the different months of the year if we were to consider them both as variable, namely:

	DEC.	JAN.	FEB.	MAR.	APR.	MAY	JUNE
m	71,000	71,000	64,000	61,000	61,000	59,000	56,000
p	∞	∞	6.4	3.8	4.7	12	46

It is seen that m decreases (the infectivity increases) throughout the epidemic and that p decreases from very large values to a minimum of 3.8 near the peak of the epidemic and then increases to very large values. (Several of the epidemics of measles in Baltimore give similar results.[6]) There is certainly little evidence that p tends to be constant, although if one considered not only the intersections of successive pairs of lines but those of all the lines, there is some tendency of the intersections to cluster around the middle of the diagram near a point corresponding to the values $p = 4.3$, $m = 61{,}000$ appropriate to the relation between total cases and peak cases; there is no point of intersection which indicates a value of p so small as 1, or even 2.

The conclusion to be drawn from this behavior of p and m during the course of the epidemic, when both are regarded as variable, seems clearly



FIGURE

The lines $\log S = \log m + \frac{1}{2p} \log K_0^1 / K_{-1}^0$ determined from the cases in successive pairs of months in Baltimore 1912–1913.

to be that it is impracticable to consider the law of mass-action as accurately representing the dynamics of an epidemic with constant values of p and m, although for some purposes it may be a highly valuable approximation. As there is no observationally well-established value of p and as the *a priori* theory leading to the use of $p = 1$ is neither particularly convincing nor well suited to reproducing such a major statistical relationship as that between total cases and peak cases, as the computed variation of the infectivity (measured in terms of $1/m$) is so sensitive to the values,[7] fixed or variable, that may be used for p, and, finally, as the changes in the infectivity which may be due to the rise and fall of the epidemic cannot at present be separated from those due to the season as such, we infer that much further study will be required before one can have any confidence in the reality of the variation of infectivity that may be computed in some particular way from the data.

[1] These PROCEEDINGS, 31, 142–147 (1945).

[2] If the case rate C is increasing (or decreasing) in geometrical progression as $C = ae^{bt}$, and if for symmetry we take C at $t + \tau/2$ and at $t - \tau/2$ and the cases K between t and $t + 2\tau$ and between $t - 2\tau$ and t we have $C(t + \tau/2) = ae^{bt}e^{b\tau/2}$, $C(t - \tau/2) = ae^{bt}e^{-b\tau/2}$ and

$$K_t^{t+2\tau} = \frac{a}{b}e^{bt}(e^{2b\tau} - 1), \quad K_{t-2\tau}^{t} = \frac{a}{b}e^{bt}(1 - e^{-bt});$$

the relation (2) is then exact. If the case rate be expanded in series,

$$C(t \pm \tau/2) = C(t) \pm \frac{\tau}{2}C'(t) + \frac{\tau^2}{8}C''(t) + \ldots,$$

$$K = 2\tau C(t) \pm 2\tau^2 C'(t) + \frac{4}{3}\tau^3\frac{C''}{2}(t) + \ldots,$$

$$\left[\frac{C(t - \tau/2)}{C(t + \tau/2)}\right]^2 = 1 - 2\frac{\tau C'(t)}{C(t)} + 2\tau^2\left[\frac{C'(t)}{C(t)}\right]^2 + \ldots, \quad \frac{K_t^{t-2\tau}}{K_{t-2\tau}^{t}} = 1 - 2\frac{\tau C'(t)}{C(t)} + 2\tau^2\left[\frac{C'(t)}{C(t)}\right]^2 + \ldots$$

and therefore the formula is good to terms in τ^2, inclusive, but inexact in the terms of order higher than τ^2 which involve higher powers of C'/C than the second and the derivatives of C or order higher than the first. As the interval τ can hardly be considered small for epidemics of measles, one may examine the accuracy of (2) for a variety of curves which may be fitted to the case rate and it appears that the formula involves a relatively small percentage error which can hardly be considered serious in view of the inevitable fluctuations in the course of observed epidemics.

[3] The point of division falls between two epidemics which had numbers of cases almost equal to the numbers of recruits during the year.

[4] The fluctuations outside the central range of epidemics are so great that for averages of only six no particular stability could be expected.

[5] The index of infectivity varies from $1 + \tanh \alpha$ to $1 - \tanh \alpha$ and with the values of α that may well arise in fitting the solution to an observed epidemic the variation may be of the order of magnitude shown in Table 1.

[6] If we write (1) as $\log m = \log S + p^{-1} (\log C_{-} - \log C)$ and differentiate with respect to the time to find the intersection of successive lines,

$$0 = \frac{1}{S}\frac{dS}{dt} - p^{-1}\left[d\frac{\log C}{dt} - \frac{d \log C}{dt}\Big|_{-}\right], \quad p = \frac{\tau d^2 \log C}{dt^2} \frac{S}{A - C}$$

where A is the rate of recruiting susceptibles to the population. Thus we should expect p to be infinite when $C = A$. The inflection points of $\log C$ would probably be well down in the trough of the seasonal curve. Thus we should expect p to be positive and near minimum at the peak of the epidemic, to increase to ∞ when $C = A$ either on the way up or on the way down, to be negative when C was between A and the presumably smaller values at which $d^2 \log C/dt^2$ becomes zero and to be positive between the two points in the trough where $d^2 \log C/dt^2 = 0$; but with actual data from an epidemic the instability due to fluctuations when $C < A$ would probably interfere with tracing the variation of the value of p. Figure 1, computed from finite differences, appears to check well enough with this general discussion based on differentiation.

[7] The values of k/m depend not only on those used for p; they depend on the general level of S and on the true number of cases (not the reported number) in so far as the variation in the value of S during the epidemic is concerned, but not in so far as the ratio K_0^1/K_{-1}^0 is concerned provided we may consider the proportion of cases reported to be constant during the epidemic. There are several hypotheses on which k/m may be computed: (1) The level of S is fixed and p is fixed but may have different values. This has been discussed above, using the level of S and the true number of cases as determined by Hedrich. (2) The value of p is fixed and unchanging but the level of S is unknown. It seems fairly clear that minor changes in the general level of S should not markedly change the variation of the index of infectivity k/m; but in some diseases on which one might wish to test the theory the general level of S might be unknown within fairly wide limits, and, as a matter of fact, some persons might differ fairly seriously with Hedrich's figures or with our conclusion that $5^1/_2$ years of recruits was a good estimate of the general level of S for measles in American cities. (3) As the statistical relation between total cases and peak cases and the theoretical expression for the period between epidemics depend effectively upon the ratio m/p rather than upon m and p severally, one might desire to discuss the changes in the "seasonal variation of infectivity" which would result from varying p and the general level m of susceptibles subject to the condition that m/p were constant. We believe that under this hypothesis the infectivity decreases throughout the major part of the epidemic but that the range of the index of infectivity diminishes as p increases.

*STRAIN SPECIFICITY AND PRODUCTION OF ANTIBIOTIC SUBSTANCES. VI. STRAIN VARIATION AND PRODUCTION OF STREPTOTHRICIN BY ACTINOMYCES LAVENDULAE**

BY SELMAN A. WAKSMAN AND ALBERT SCHATZ

NEW JERSEY AGRICULTURAL EXPERIMENT STATION, RUTGERS UNIVERSITY

Communicated June 13, 1945

Attention has already been directed[4, 10] to the fact that actinomycetes as a whole, and certain groups or species in particular, vary greatly in their cultural, physiological and morphological properties. The variable char-

acteristics include the nature and relative abundance of the aerial mycelium, the formation of various soluble and insoluble pigments, and the capacity of producing one or more antibiotic substances. These variations are obtained both for strains of presumably the same organism isolated from different substrates, and for variants of a single culture when grown on a variety of media and under different conditions. Moreover, the same culture frequently exhibits variations when grown on a certain medium for a considerable length of time. Some characteristic cultural or morphological properties may thus be gained or lost.

Schaal[8] reported that isolates of *Actinomyces scabies*, the causative agent of potato scab, produce variants that are culturally and morphologically different from the parent strains and from one another; they also differ in their pathogenicity upon potatoes. Some of these variants arise spontaneously, as previously demonstrated by Jensen[3] and others.[7] On the other hand, some investigators[2, 6, 11] have not accepted the concept of a fundamental change in the characteristics of the organism which would render it a distinct species; they believe that the variations are only temporary in nature and that the characteristics of the organisms remain constant provided cultural and environmental conditions are constant.

The phenomenon of variation among the actinomycetes has been a major cause of confusion in the nomenclature and classification of these organisms. A species once described on the basis of its cultural or morphological properties may not necessarily be recognized from the original description after several years of culture.[4] This may be particularly true of organisms isolated from different substrates and in different laboratories, especially if those organisms vary in some one cultural characteristic. The same species isolated by different workers has often been looked upon as representing a totally different oganism, especially when no sufficient recognition was given to the potential variability of the organisms. This fact often accounts for the different designations attached to the same species. An organism described by Müller,[5] in 1908, as *Actinomyces coelicolor* was later designated as *A. violaceus-ruber*,[12, 16] *A. tricolor* and *A. violaceus;* however, Conn,[1] in a study of the chemical nature of the pigment produced by different cultures of this organism, was led to suggest that one may be dealing, after all, with distinctly different species. The pigment as well may be a variable character.

Among the various actinomycetes belonging to the genus *Streptomyces*[17] formerly designated as *Actinomyces*, Waksman and Curtis[16] described, in 1916, an organism under the name of *A. lavendulae*. When freshly isolated, the culture of this organism was characterized by the production of a lavender-colored aerial mycelium on synthetic media and of a soluble brown pigment on protein media. This culture was not tested, when first isolated, for its capacity to form antibiotic substances, but when tested

recently, namely, after nearly 30 years in culture, it was found to show only weak antibiotic activity. This need not, of course, indicate that originally the culture may not have been an active producer of some antibiotic substance, since this property, as well, may have been either lost or largely reduced under conditions of artificial culture for this long period of time.

Since the first description of *A. lavendulae*, other strains of this organism or forms closely related to it have been isolated from the soil by the senior author and by various students and collaborators working in his laboratory.[16, 19, 18] Several strains were also received from other laboratories in this country. Many of these strains differed from the description of the original culture in some of the cultural characteristics, such as intensity or shade of pigmentation of the aerial mycelium and nature of soluble pigment when grown on organic media. However, these differences were considered to be only minor or quantitative in nature, and not of sufficient significance to justify the designation of any of the new strains as different species. This point of view seemed justified since the original culture itself underwent, during the long period of cultivation, marked changes in pigmentation, and abundance and nature of aerial mycelium.

Among the various isolations of *A. lavendulae* cultures in our laboratory, there was one that possessed† distinct antibiotic properties. The isolation of this culture happened to coincide with an intense interest in the laboratory[18] in the subject of antibiotic substances. On further study,[19] this culture proved to have the capacity of producing an active antibiotic agent which was designated as streptothricin. However, it was found to be rather variable in nature, as shown by the fact that a number of variants, differing in their capacity to produce streptothricin, could be readily isolated[13] from it. These variants were found, more recently, to differ also in certain other cultural and physiological characteristics, notably the production and nature of the soluble pigment, and the presence of aerial mycelium.

Experimental.—The degrees of cultural, physiological and morphological variation among different strains of *A. lavendulae* and of closely related forms can be illustrated even in a comparison of only several cultures. Among the physiological properties of the strains, emphasis was laid upon the production of the antibiotic substance streptothricin. These cultures are listed here under the various laboratory numbers.

No. 3330. The original strain of *A. lavendulae* which was isolated in 1915, and has since been maintained in the culture collection of the Department of Microbiology of the New Jersey Agricultural Experiment Station.

No. 3440-8. A strain of the active streptothricin-producing culture of *A. lavendulae*, isolated in 1941. Among many strains, this and the following one were selected for this study.

TABLE 1

CULTURAL AND MORPHOLOGICAL CHARACTERISTICS OF DIFFERENT STRAINS OF *A. lavendulae*

ORGANISM	NUTRIENT AGAR		GELATIN		MILK		POTATO PLUG		DEXTROSE ASPARAGINE AGAR	
	AM[a]	SP	LIQ.	SP	COAGU-LATION	PIG-MENT	AM	SP	AM	SPORULATING HYPHAE
A. lavendulae[b]	O	Br	+	O-Br	O	Br	O	Black	Lavender	Spirals and straight
No. 3330[c]	White	None	+++	O	O	Br	O	O	White	
No. 3440-8	O	Br	++	Br	O	Br	O	Black	Lavender to rose	Compact spirals
No. 3440-14	O	Br	++	Br	O	Br	O	Black	Lavender to rose	Straight
No. 3445	O	Br	+++	Br	O	Br	O	Black	Lavender to rose	Straight, some spirals
No. 3465	White	Br	+	Br	O	Br	O	Black	Lavender to rose	Compact spirals

[a] AM = aerial mycelium; SP = soluble pigment; Br = brown; Liq. = liquefaction.
[b] Based on original description of organisms.
[c] Original culture of *A. lavendulae*.

No. 3440-14. Another strain of the streptothricin-producing culture of *A. lavendulae.*

No. 3445. A culture isolated from the soil.

No. 3465. A culture isolated in our laboratory in 1943 from a washed agar plate with *Mycobacterium tuberculosis* as the source of nitrogen. It is active, as determined by the streak method, against gram-negative bacteria and certain mycobacteria, but not against the gram-positive bacteria. Culture filtrates obtained by growing the organism on different media have been found capable of inhibiting only certain mycobacteria.

A brief summary of the cultural and morphological characteristics of these strains is presented in table 1. For comparative purposes, some characteristics of *A. lavendulae* as reported in the original descriptions,[12] are also included. Strains 8 and 14, derived from the same parent culture (No. 3440), were identical in most of their cultural characteristics. They differed, however, in three important properties not reported in the table. Number 8 produced a brown diffusible pigment on glucose-peptone agar after 2 weeks, whereas No. 14 did not; the latter produced much more streptothricin than the former; No. 14 grew normally on different media, in a manner comparable to the original culture, whereas No. 8 produced after 24 hours, on glucose-peptone slants, a distinct blue diffusible pigment which gradually became brown on the second or third day of incubation.

The results of comparative studies on the production of an antibiotic

substance by the various cultures of *A. lavendulae*, grown under stationary and submerged conditions, are given in table 2. Strains No. 3330 and No. 3465 were completely inactive under these conditions, whereas the other three strains showed varying degrees of activity. The nature of the antibiotic action was of distinct interest. All strains were either completely or nearly inactive against *Bacillus mycoides;* they were all active against both *Escherichia coli* and *Proteus vulgaris*, as well as against *B. subtilis;* the activity against the last was always greater than against *E. coli*. These specific antibacterial properties tend to indicate that the antibiotic substances produced by the various strains is of the streptothricin type.[18, 19]

Certain interesting differences in the nature of the antibiotic activity of strains No. 3440-8 and No. 3440-14 on the one hand, and of strain No. 3445 on the other, are also to be noted. Whereas the antibiotic substance produced by the first two strains was alike in all respects, save a difference in quantitative production, the substance formed by No. 3445 showed also certain differences that may be qualitative in nature. First, this substance produced in stationary culture had limited action against *B. mycoides;* second, the ratio of the antibiotic activity against *B. subtilis* as compared to *E. coli* was usually wider for strain No. 3445. Further information, however, tends to indicate that possibly this difference in nature of the antibiotic substance produced by the two cultures may only be quantitative rather than qualitative, as brought out in table 3.

Strain No. 3440-8 was found to undergo considerable variation on further cultivation. Some of the variants obtained from this strain produced on glucose-peptone a blue diffusible pigment; others formed a brown pigment The variants producing blue pigment had a pale blue vegetative mycelium with scattered, small pin-point areas of deep blue. Upon complete sporulation, the vegetative growth was covered with thick lavender-colored aerial mycelium; occasional sunken areas were of a somewhat slightly bluish tinge, these areas corresponding to the pin-point regions of the deeper blue. The under surface of the vegetative growth was cream colored except for the small blue spots. The other variants produced a colorless to cream-colored vegetative growth free of any blue pigment whatsoever; one to two days later, a brown diffusible pigment appeared, the growth becoming covered with abundant lavender-colored mycelium. On subsequent transfer on fungus agar slants, the two types of variants proved to be rather stable.

Two distinct variants were thus obtained from strain 3440-8: (*a*) blue vegetative growth, a diffusible pigment which was initially blue, and lavender aerial mycelium of a slight blue tinge; (*b*) cream-colored vegetative growth, soluble brown pigment in peptone media, and lavender aerial mycelium. Two variants were also isolated from sectors of colonies of

TABLE 2

PRODUCTION OF AN ANTIBIOTIC SUBSTANCE BY DIFFERENT STRAINS OF *A. lavendula*

CULTURE NO.	DILUTION UNITS/ML. OF FILTRATE *E. coli*	*B. mycoides*	*Pr. vulgaris*	*B. subtilis*	CUP TEST, UNITS[a]
	Shaken cultures, 2 days' incubation				
3330	0	0	0	0	0
3440-8	0	0	0	30	3
3440-14	75	0	50	100	22
3445	25	0	10	150	15
3465	0	0	0	0	0
	Stationary cultures, 7 days' incubation				
3330	0	0	0	0	0
3440-8	0	0	10	75	7
3440-14	25	0	30	200	34
3445	25	10	30	300	24
3465	0	0	0	0	0

[a] Filtrates tested by cup method with *B. subtilis* as test organism.

TABLE 3

PRODUCTION OF AN ANTIBIOTIC SUBSTANCE BY *A. lavendulae* 3440-14 AND ACTINOMYCES 3445

Starch-Tryptone Medium, Shaken Culture Agar Streak Method, Activity Units

INCUBATION, DAYS	CULTURE NO.	*E. coli*	*B. mycoides*	*B. subtilis*	*S. aureus*	CUP METHOD, MM. ZONE
2	3440-14	50	<10	>300	50	33.5
2	3445	150	20	>300	150	31.0
3	3440-14	100	5	500	100	25.5
3	3445	100	20	750	150	25.5
4	3440-14	100	<10	300	75	..
4	3445	100	5	300	100	..

TABLE 4

PRODUCTION OF STREPTOTHRICIN BY TWO STRAIN OF *A. lavendulae* AND TWO VARIANTS[a]

Shaken Cultures

CULTURE NO.	UNITS/ML. OF FILTRATE, AFTER DAYS 3	4	6
3440-8	38	25	26
3440-8 (*a*)[a]	<2	<2	Trace
3440-8 (*b*)[a]	64	50	42
3440-14	40	23	22
3440-14 (*a*)	23	23	..
3440-14 (*b*)	0	0	0

[a] (*a*) and (*b*) are variants obtained from strains Nos. 3440-8 and 3440-14.

culture 3440-14 grown on agar media; (*a*) white aerial mycelium, sometimes showing a faint shade of pink; and (*b*) devoid of aerial mycelium, except for a scant growth of sporulating aerial hyphae on some very old slants.

These four variants were compared with the original cultures No. 3440-8 and No. 3440-14, for their ability to produce streptothricin in shaken cultures. The results are presented in table 4. Strain 3440-8 was occasionally found to be less active than No. 3440-14. This may be due to

the fact that strain 3440-8 tends to separate into the two variants (*a*) and (*b*), the first of which is almost completely inactive, whereas the second is more active than the parent strain. In the case of No. 3440-14, however, the white sporulating variant (*a*) was active, although somewhat less so than the lavender parent strain; the (*b*) strain, free of aerial mycelium, was completely inactive.

In the case of strain 3440-14, therefore, the production of the antibiotic agent was associated with the ability of the culture to form aerial mycelium. This is similar to the results obtained in strain variation studies of *A. griseus*.[10] In the case of both organisms the variants which failed to produce aerial mycelium likewise produced inactive culture filtrates. This statement must apply, of course, only to the inactive asporogenous strains and their respective active parent cultures, since other independent isolations of both *A. griseus* and *A. lavendulae* have been found completely inactive in spite of the fact that they produced aerial mycelium.

Summary.—Different cultures of *A. lavendulae* differ greatly in their capacity to produce the antibiotic substance streptothricin.

Active cultures of *A. lavendulae* were found to give variants that differed morphologically, culturally and physiologically from the parent strain. Variants free of aerial mycelium did not produce any streptothricin.

* Journal Series Paper, New Jersey Agricultural Experiment Station, Rutgers University, Department of Microbiology.

† This culture was isolated by Dr. W. Kocholaty, of the University of Pennsylvania, while working in our laboratory.

[1] Conn, J., *Jour. Bact.*, **46**, 133–149 (1943).

[2] Duché, J., *Les Actinomyces du groupe albus*, P. Lechevalier, Paris, 1934.

[3] Jensen, H. L., *Proc. Linn. Soc. N. S. Wales*, **56**, 79–98 (1931).

[4] Kriss, A. E., "Variability of Actinomycetes," *Acad. Sci., U.S.S.R. Moscow*, 101 pp. (1937).

[5] Müller, R., *Zentrbl. Bakt. Parasit.*, I, **46**, 195–212 (1908).

[6] Plotho, O. V., *Archiv Mikrob.*, **11**, 33–72 (1940).

[7] Rippel, A., and Witter, P., *Ibid.*, **5**, 24–30 (1934).

[8] Schaal, L. A., *Jour. Agr. Res.*, **69**, 169–186 (1944).

[9] Schatz, A., Bugie, E., and Waksman, S. A., *Proc. Soc. Exp. Biol. Med.*, **55**, 66–69 (1944).

[10] Schatz, A., and Waksman, S. A., *Proc. Nat. Acad. Sci.*, **31**, 129–137 (1945).

[11] Tempel, E., *Archiv Mikrob.*, **2**, 40–71 (1931).

[12] Waksman, S. A., *Soil Science*, **8**, 71–215 (1919).

[13] Waksman, S. A., *Jour. Bact.*, **46**, 299–310 (1943).

[14] Waksman, S. A., Bugie, E., and Schatz, A., *Proc. Staff Meet. Mayo Clinic*, **19**, 537–548 (1944).

[15] Waksman, S. A., and Curtis, R. E., *Soil Sci.*, **1**, 99–134 (1916).

[16] Waksman, S. A., and Curtis, R. E., *Ibid.*, **6**, 309–319 (1918).

[17] Waksman, S. A., and Henrici, A. T., *Jour. Bact.*, **46**, 337–341 (1943).

[18] Waksman, S. A., Horning, E. S., Welsch, M., and Woodruff, H. B., *Soil Sci.*, **54**, 281–296 (1942).

[19] Waksman, S. A., and Woodruff, H. B., *Proc. Soc. Exp. Biol. Med.*, **49**, 207–210 (1942).

PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 31 August 15, 1945 Number 8

*X-RAY INDUCED MUTANT STRAINS OF ESCHERICHIA COLI**

BY E. L. TATUM

SCHOOL OF BIOLOGICAL SCIENCES, STANFORD UNIVERSITY

Communicated June 20, 1945

In a previous paper the production by means of x-ray treatment of a number of growth-factors requiring mutant strains of *Escherichia coli* and of *Acetobacter melanogenum* was described.[1] These results were interpreted as indicating an analogy with the induction by irradiation of true gene mutations in *Neurospora* with similar biochemical changes.[2, 3, 4] Comparable changes have been induced in other strains of *E. coli*.[5, 6] The present work was undertaken with the aims of producing more "biochemical" mutant strains of *E. coli* for biochemical study, of providing convincing evidence of derivation of these strains and of comparing the mutation rates of irradiated and control cultures. A considerable number of new mutant strains have been obtained after further irradiation of two biochemical mutant strains previously described.[1] These new strains require growth-factors in addition to those required by the parent strains. The double requirements of the new strains thus provide reasonable certainty of their derivation from their parent strains. The irradiation significantly raised the mutation rate from one in 2000 isolations to 16 in the same number of isolations. Furthermore, the biochemical specificities of those mutant strains which have been analyzed are consistent with the view that each differs from the original stock in only one biochemical reaction. The evidence therefore supports the view that treatment with x-rays produces heritable defects in synthetic reactions in bacteria, biologically and biochemically analogous to those resulting from single gene mutations in *Neurospora*. This is presumptive evidence for the existence of genes in bacteria, perhaps contained in the nuclear structures which have been observed in a number of bacteria.[7]

Experimental.—Two previously obtained mutant strains of *E. coli* have been used in these investigations. Strain 58 is characterized by requiring biotin, and strain 679, threonine.[1] These mutant strains were used, rather than the wild type strain from which both were derived, so that their char-

acteristic growth-factor requirements would serve as biochemical "markers." Both of these strains had been maintained for over a year on complete agar medium (containing yeast extract, peptone and glucose) with no changes in their minimum requirements. Twenty-four-hour-old cultures grown in complete liquid medium with shaking were irradiated with approximately 180,000 r units during a period of 45 minutes, as previously described.[1] After a further incubation period of four hours, samples of the irradiated suspensions were plated out in complete medium, incubated for several days and then isolated colonies transferred to complete agar medium. These isolated strains were then tested for growth in the liquid minimal medium[1] which for testing new strains derived from strain 58 was supplemented with 0.1 μg. biotin per 100 ml., and for testing those derived from strain 679 was supplemented with 10 mg. *dl*-threonine per 100 ml. Cultures which consistently failed to grow in the minimal media were then tested for their specific additional growth requirements. The data in table 1 summarize the results obtained. A mutation rate of around 1 per cent was obtained with both irradiated strains, as contrasted with a rate of 0.05 with the unirradiated culture, apparently a significant increase on irradiation to approximately the same value previously obtained with *Acetobacter melanogenum*, but somewhat higher than that obtained with the original strain of *E. coli* (K-12).[1]

TABLE 1

EFFECT OF X-RAY TREATMENT ON PRODUCTION OF MUTANT STRAINS OF *E. coli*

STRAIN OF ORIGIN[1]	X-RAY TREATMENT	VIABLE CELLS, %*	NUMBER OF COLONIES ISOLATED	DESIGNATION OF STRAINS	NUMBER OF MUTANT STRAINS
58	180,000 r	0.001	941	58-1 to 58-941	9
679	180,000 r	0.01	800	679-1 to 679-800	7
58	None	100.0†	1902	58-1001 to 58-2902	1

* Calculated from direct microscopic and plate counts.

† This culture was 72 hours old at the time of plating.

Of the 17 mutants obtained all but three have been found to require the addition of only one known growth factor for normal growth. Table 2 gives the substance required by each of these strains, together with the concentration which is required for $^1/_2$ maximal growth from turbidity measurements. The growth response of each of these strains is a function of the concentration of the growth factor supplied. Most of the mutant strains showed amino acid deficiencies, but three strains required thiamin, and three failed to grow on hydrolyzed casein or on a mixture of vitamins, but grew well on yeast extract. The only apparent duplications of requirements were those of the proline requiring strains 679-183 and 679-440, and of the thiamin requiring strains 58-593 and 58-610. In all other cases, either different substances were required or the strain of origin was different,

so that of the 17 mutant strains isolated, 14 were definitely of independent origin.

After the required substance was determined, each strain derived from strain 58 was re-tested for its requirement for biotin, and each one derived from strain 679 re-tested for its need for threonine. In every case the original requirement persisted in addition to the second induced requirement. The possession of these double requirements is definite evidence of the origin of the new strains. The biochemical specificities of some of these mutant strains have also been tested. Table 2 also lists the biochemically related substances tested for activity on each mutant strain.

TABLE 2

DESCRIPTION OF MUTANT STRAINS OF *E. coli*

STRAIN NUMBER	SUBSTANCE REQUIRED	CONCENTRATION GIVING 1/2 MAXIMAL GROWTH IN 72 HOURS* PER 10 ML.	RELATED SUBSTANCES TESTED	
			ACTIVE	INACTIVE
58-161	Methionine	0.01 mg.	Homocystine	Na_2S, cystine
58-278	Phenylalanine	0.016 mg.	....	Tyrosine
58-309	Cystine	0.16 mg.	....	Na_2S, methionine ± serine, homocystine ± serine
58-336	Isoleucine	0.04 mg.	Hydroxy- and keto-acid analogues	....
58-580	Thiamin	0.004 μg.	....	Thiazole + pyrimidine
58-593	Thiamin	0.003 μg.	Thiazole	Pyrimidine
58-610	Thiamin	0.003 μg.	Thiazole	Pyrimidine
58-741	Histidine	0.003 mg.	....	....
679-183	Proline	0.02 mg.	....	Ornithine, arginine, glutamic acid, hydroxyproline
679-440	Proline	0.03 mg.	....	Ornithine, arginine, glutamic acid, hydroxyproline
679-662	Glutamic acid or proline†	0.05 mg. 0.05 mg.	....	Ornithine, arginine, hydroxyproline
679-680	Leucine	0.03 mg.	Hydroxy-acid analogue	....
58-2651‡	Proline	0.03 mg.	....	Ornithine, arginine, glutamic acid, hydroxyproline
58-178 679-447 679-455	Yeast extract	....	....	Vitamins, hydrolyzed casein

* From growth curves obtained with an Evelyn photoelectric colorimeter.

† The growth of this strain on proline is slower than on glutamic acid, although quantitatively similar responses are obtained in 72 hours.

‡ Obtained from unirradiated material.

Strains 58-336 and 679-680, which require isoleucine and leucine respectively, can grow with the corresponding hydroxy-acid analogues, and in the one case tested (strain 58-336) also on the keto-acid analogue. Presumably the syntheses of these amino acids are blocked at stages in the formation of the carbon skeletons prior to the formation of the keto acids, which can be aminated to form the amino acids. The methionine requiring strain 58-161 presumably cannot form methionine from reduced sulfur (Na_2S) or from cystine, but can do so to a limited extent from homocystine, a process which would require the methylation of homocysteine. The cystine requirement of strain 58-309 is extremely specific. Apparently cystine cannot be formed from Na_2S, nor from methionine or homocystine even in the presence of added serine. If the normal synthesis of cysteine in *E. coli* involves the transfer of the —SH group from homocysteine to serine by the reaction studied by Binkley and du Vigneaud,[8] this strain must be unable to transfer the sulfhydryl group to serine.

None of the three strains which respond to proline use ornithine, arginine or hydroxy-proline. However, one can grow on either proline or on glutamic acid. Proline and glutamic acid are apparently interconvertible in strain 679-662, but not in the other two strains. If the failure of synthesis of proline in each of these strains results from the failure of a single biochemical reaction, these facts suggest that ornithine is not involved in the synthesis of proline by *E. coli*, in which glutamic acid may be the immediate precursor of proline.[9]

There are two types of thiamin requiring strains, both analogous to known mutant strains of *N. crassa*.[4, 10] The two representatives of one type, strains 58-593 and 58-610, apparently are unable to synthesize the vitamin thiazole, but can form the pyrimidine component and couple the two components. Strain 58-580, however, cannot carry out this coupling, since it requires intact thiamin.

Summary.—Following a second x-ray treatment of two mutant strains of *Escherichia coli* 16 new mutant strains have been obtained from 1741 single colony isolations. One mutant strain was obtained from 1902 isolations from unirradiated material.

Fourteen of these strains have been found to require single known growth factors in addition to the requirement of the strain from which they were derived for biotin or for threonine. The growth factors required by these strains include proline, glutamic acid or proline, leucine, isoleucine, phenylalanine, methionine, cystine, histidine, thiamin and vitamin thiazole. Three strains require unidentified substances present in yeast extract.

These results are consistent with the conclusion that growth-factor requirements in bacteria result from heritable changes analogous to true gene mutations.

* Work supported in part by a grant from the Rockefeller Foundation. The author wishes to thank Dr. David Bonner for supplying the hydroxy- and keto-acid analogues of leucine and isoleucine, and Miss Esther M. Zimmer for technical assistance in carrying out this investigation.

[1] Gray, C. H., and Tatum, E. L., these PROCEEDINGS, **30**, 404 (1944).

[2] Beadle, G. W., and Tatum, E. L., *Ibid.*, **27**, 499 (1941).

[3] Horowitz, N. H., Bonner, D., Mitchell, H. K., Tatum, E. L., and Beadle, G. W., *Amer. Nat.* (in press).

[4] Tatum, E. L., and Beadle, G. W., *Annals Mo. Botanical Garden*, **32**, 125 (1945).

[5] Roepke, R. R., Libby, R. L., and Small, M. H., *Jour. Bact.*, **48**, 401 (1944).

[6] Anderson, E. H., these PROCEEDINGS, **30**, 397 (1944).

[7] For descriptions and for further references see Robinow, C. F., *Jour. Hygiene*, **43**, 413 (1944).

[8] Binkley, F., and du Vigneaud, V., *Jour. Biol. Chem.*, **144**, 507 (1942).

[9] For a discussion of these relationships and for further references, see Shemin, D., and Rittenberg, D., *Jour. Biol. Chem.*, **158**, 71 (1945).

[10] Tatum, E. L., and Bell, T. T. (in preparation).

CARBON DIOXIDE UTILIZATION IN THE SYNTHESIS OF ACETIC ACID BY CLOSTRIDIUM THERMOACETICUM

BY H. A. BARKER AND M. D. KAMEN*

DIVISION OF PLANT NUTRITION, UNIVERSITY OF CALIFORNIA

Communicated June 20, 1945

In most carbohydrate fermentations the appearance of acetic acid or other C_2 products is accompanied by at least an equimolar quantity of carbon dioxide or other C_1 product. The carbohydrate fermentations caused by *Clostridium thermoaceticum* are a noteworthy exception to this generalization since the only product other than cell material is acetic acid.[1, 2] Specifically, glucose and xylose are acted upon by this organism to yield, respectively, about 2.65 and 2.25 moles of acetic acid per mole of carbohydrate decomposed. When pyruvate is the substrate, both acetic acid and carbon dioxide are formed but the yield of the latter is low, corresponding to about 0.5 mole per mole of pyruvate.

To account for the high yield of acetic acid and the low yield of carbon dioxide from all three substrates, it has been postulated that *Clostridium thermoaceticum*, like *Clostridium aceticum*[3, 4] and *Clostridium* acidi-urici,[5] uses carbon dioxide as an oxidizing agent in such a way that it is condensed with a second molecule and reduced to acetic acid. If this occurs in the decomposition of glucose, for example, at least two alternative mechanisms, illustrated by the following sets of equations, are possible.

I. $$C_6H_{12}O_6 + 2H_2O = 2CH_3COOH + 2CO_2 + 8H \quad (1)$$

$$8H + 2CO_2 = CH_3COOH + 2H_2O \quad (2)$$

II. $$C_6H_{12}O_6 + 6H_2O = 6CO_2 + 24H \quad (3)$$

$$6CO_2 + 24H = 3CH_3COOH \quad (4)$$

For either scheme the over-all reaction is

$$C_6H_{12}O_6 = 3CH_3COOH. \quad (5)$$

Scheme I represents the fermentation as a partial oxidation of glucose to acetic acid and carbon dioxide (equation 1) accompanied by the reduction of 2 moles of carbon dioxide to acetic acid (equation 2). The oxidation of glucose presumably follows the usual glycolytic mechanism involving pyruvic acid. Scheme II represents the fermentation as a complete oxidation of glucose (equation 3) accompanied by the reduction of 6 moles of carbon dioxide (equation 4) to acetic acid. Although Scheme II appears less likely to represent the true course of the fermentation than does Scheme I, it must nevertheless be considered as a possibility.

In this paper experiments will be described which provide direct evidence for the conversion of carbon dioxide to acetic acid and for the decomposition of glucose in accordance with Scheme I. This evidence has been obtained[6] by the use of carbon dioxide labeled with the long-lived carbon isotope, C^{14}.

Experimental Methods.—For the detection of radiation emitted by C^{14} a Geiger counter of "bell-jar" type with a thin mica window was used. The mica window had a thickness corresponding to about 3.5 mg. per $cm.^2$ and an area of approximately 25 $cm.^2$ It was supported by a brass grid cut to allow maximal passage of the radiation. Samples consisting of barium carbonate or barium acetate were spread thinly and evenly on duraluminum discs and dried at 100°C. These were placed about 1 mm. below the window in a standard fixed position. Variations in sensitivity of the counter were corrected for by reference to a standard sample of barium carbonate prepared in the same way. Corrections for self-absorption, necessitated by the softness of the beta rays (0.15 M. E. V.) emitted, were determined from a curve constructed from counts on standard samples of known activity and varying thickness.† The long half-life of C^{14} ($2.5 \pm 1 \times 10^4$ yrs.)[7] obviated corrections for decay. The counting circuit was of a conventional type and requires no description.

The bacteria were grown in a medium of the following composition in g. per 100 ml.: glucose 0.15–0.7, tryptone 0.5, yeast extract 0.5, pH 6.6, phosphate buffer 0.9, $(NH_4)_2SO_4$ 0.05, $MgSO_4 \cdot 7H_2O$ 0.01, sodium thioglycollate 0.05, $Na_2C^*O_3$ about 0.2 (* indicates carbon labeled with C^{14}). The phosphate buffer and labeled sodium carbonate were added after autoclaving. In a typical experiment the total volume of medium was about 12 ml. Oxygen was removed by means of an Oxsorbent seal and

the culture tube was closed with a ground-glass stopper to prevent loss of carbon dioxide. After incubating for 3 to 6 days at 55°C. cultures were analyzed and the C^{14} content of the carbon dioxide and the fermentation products was determined.

Results.—In table 1 are presented data from an experiment such as that described above. The glucose concentration was 0.7 g. per 100 ml.

TABLE 1

THE FERMENTATION OF GLUCOSE IN THE PRESENCE OF C^*O_2

Experiment 1

SUBSTANCE	MG./10 ML.	CTS./MIN./MG.	TOTAL CTS./MIN.
Glucose fermented	54.0	...	....
Initial carbon dioxide as $BaCO_3$	(22.3)	117	2610 ± 50
Final carbon dioxide as $BaCO_3$	22.3	5.7	128 ± 20
Acetic acid formed as $BaAc_2$	101.6	19.9	2020 ± 40
Cell material (residue from T. C. A. extraction)	2.5	12.8	32 ± 4
Trichloracetic acid extract of cells	4.5	1.5	7 ± 4
Non-volatile cell-free material	...	...	96 ± 30
Total C^{14} activity in products			2155
C^{14} recovered in per cent			2155/2482 × 100 = 87

It can be seen that most (94%) of the C^{14} disappeared from the added carbon dioxide during the fermentation and a large part of it (81%) was recovered in acetic acid. Adequate evidence that the C^{14} was actually present in acetic acid rather than some associated compound was provided by distilling the volatile acid, using a modified Duclaux method, and determining the specific activity of the barium acetate derived from successive distillation fractions. The specific activity was found to be the same in all fractions within the limits of experimental accuracy (table 2).

TABLE 2

SPECIFIC ACTIVITIES OF DUCLAUX FRACTIONS OF ACETIC ACID

Experiment 1 (total volume 110 ml.)

DUCLAUX FRACTION	CTS./MIN./MG. $BaAc_2$
0– 40 ml.	17.9
40– 80 ml.	18.9
80–110 ml.	18.7

A small part of the C^{14} was also present in the cells after removal of acetic acid and carbon dioxide, and in the non-volatile fraction of the medium after removal of the cells. It is noteworthy that the quantity of C^{14} per mg. of dry cells was of the same order of magnitude as in the barium acetate. It seems reasonable to conclude that a considerable part of the cell material was synthesized from carbon dioxide.

Only about one-fifth of the C^{14} in the cells could be extracted with 4% trichloracetic acid; the remainder must have been present in proteins,

lipoids and other acid-insoluble forms. The small yield of acid-extractable C^{14} was due possibly to the relatively long incubation which was continued for several days after growth ceased. The bacterial cells may have autolyzed, releasing soluble constituents. This could account for the considerable amount of non-volatile C^{14} present in the medium after removal of the bacteria. The actual weight of this fraction could not be estimated because of the large amount of inorganic salt present; for the same reason the C^{14} content of this fraction could be only roughly determined.

To determine the distribution of C^{14} in the acetic acid, its barium salt was first decarboxylated according to the equation

$$(CH_3COO)_2Ba = CH_3COCH_3 + BaCO_3.$$

In control experiments with synthetic CH_3C^*OOH, the barium carbonate was shown to originate from the carboxyl carbon of acetic acid (table 3, column 2). Hence, this method was used to estimate the C^{14} content of the carboxyl carbon. To determine the C^{14} content of the methyl group, the acetone derived from the decarboxylation was oxidized with alkaline iodine to iodoform and acetic acid. The iodoform, which was shown to originate from the methyl group of acetic acid (table 3, column 2), was oxidized to carbon dioxide and converted to barium carbonate for counting C^{14}.

TABLE 3

DISTRIBUTION OF C^{14} IN SYNTHETIC AND FERMENTATION ACETIC ACID

(The figures give the percentage of the total C^{14} in each atom)

		FERMENTATION ACETIC ACID	
CARBON ATOM	SYNTHETIC CH_3C^*OOH	EXPT. 1	EXPT. 3
Methyl	0 ± 3	49 ± 2	40 ± 2
Carboxyl	100 ± 2	51 ± 2	60 ± 2

Data from two experiments are given in table 3, columns 3 and 4. It can be seen that in experiment 1 the C^{14} in the acetic acid produced by fermentation was almost equally distributed between the methyl and carboxyl groups as would be expected if a total synthesis of acetic acid from carbon dioxide had occurred. In experiment 3, a different result was obtained; about 50% more C^{14} was present in the carboxyl than in the methyl group. The difference in C^{14} distribution between the two experiments appears to be significant but we have no explanation for it. Further work will be necessary to elucidate the relation between the experimental conditions and the isotope distribution in acetic acid. There is, however, a possible explanation for the fact that in experiment 3 the carboxyl group had a higher C^{14} content than the methyl group. The synthesis of acetic acid from carbon dioxide by *Cl. thermoaceticum* may occur in two ways, one involving the fixation of carbon dioxide in both the methyl and carboxyl groups, the other involving only fixation in the carboxyl group.

The relative rates of the two processes would then determine the distribution of the isotope. A fixation of carbon dioxide exclusively in the carboxyl group of acetic acid has been observed by Slade, *et al.*,[5] using *Aerobacter indologenes* and *Clostridium welchii*.

Having established the conversion of carbon dioxide to acetic acid, it seemed desirable to find out how much carbon dioxide was produced as an intermediate in the glucose fermentation. This can be calculated from a knowledge of the C^{14} contents of the initial and final carbon dioxide if it is assumed that the formation of carbon dioxide from sugar is the only process which causes dilution of the labeled carbon dioxide. One specific reaction which must not occur is an exchange of C^{14} between carbon dioxide and preformed acetic acid.

The evidence against a reversible exchange of C^{14} between carbon dioxide and acetic acid is of two types. First, a wide variation in the molar ratio of the C^{14} contents of acetic acid and carbon dioxide was observed in different experiments. The range was from 0.95 (experiment 2, table 4) to 2.26 (experiment 1, table 1). A rapid exchange of C^{14} between the two compounds should result in a constant ratio. Secondly, an experiment was performed in which glucose was fermented in the presence of acetate, labeled in both the methyl and carboxyl positions, and ordinary carbon dioxide. At the beginning of the experiment 0.215 mM of acetic acid giving 726 $\pm$ 15 cts./min. was present per 10 ml. of medium. After the fermentation 0.385 mM of acetic acid giving 706 $\pm$ 15 cts./min. was recovered. The final carbon dioxide (0.049 mM) gave only 9.8 $\pm$ 1 cts./min., or about 1% of the C^{14} initially added as acetate. From this result it must be concluded that whereas a conversion of acetic acid carbon to carbon dioxide does occur the rate of the reaction is so slow as to be negligible under the conditions of these experiments.

To calculate the intermediate carbon dioxide production we shall assume then that the only pertinent reactions are the following:

$$(a)\quad \text{glucose } (C^{12}) \rightarrow CO_2\ (C^{12})$$

$$(b)\quad CO_2\ (C^{12} + C^{14}) \rightarrow \text{acetic acid } (C^{12} + C^{14}).$$

The carbon dioxide will always consist of a mixture of $C^{14}O_2$ and $C^{12}O_2$. Let x be the quantity of $C^{14}O_2$ per unit volume at any time during the fermentation and let x_0 be the initial, and x_f the final $C^{14}O_2$. Further, let V represent the amount of $C^{12}O_2$ plus $C^{14}O_2$ converted to acetic acid at any time; V also equals the amount of $C^{12}O_2$ formed from glucose since there is no net change in carbon dioxide. Va is the constant amount of carbon dioxide present throughout the fermentation and V_f is the total carbon dioxide formed or utilized during the experiment.

When a small quantity (ΔV) of $C^{14}O_2 + C^{12}O_2$ is converted to acetic

acid and an equal amount of $C^{12}O_2$ is formed from glucose, the decrease in $C^{14}O_2$ $(-\Delta x)$ is given by the expression

$$-\Delta x = \frac{\Delta V}{Va + \Delta V} \cdot x$$

which means that the $C^{14}O_2$ (x) is decreased by a fraction equal to the carbon dioxide used (ΔV) divided by the total carbon dioxide present $(Va + \Delta V)$. Dividing by x we get

$$\frac{-\Delta x}{x} = \frac{\Delta V}{Va + \Delta V}.$$

In the limit as ΔV is decreased

$$\frac{-dx}{x} = \frac{dV}{Va + dV} = \frac{dV}{Va}.$$

Integrating this expression between the limits x_0 and x_f for x, and 0 and V_f for V, we find

$$-\int_{x_0}^{x_f} \frac{dx}{x_1} = \ln \frac{x_0}{x_f} = 1/Va \int_0^{V_f} dV = \frac{V_f}{Va}$$

or

$$V_f = 2.3 \cdot Va \log x_0/x_f.$$

V_f has the same units as Va. V_f must be divided by the quantity of glucose fermented (= moles of acetic acid formed divided by 2.65) to give carbon dioxide production per unit of glucose.

When this method of calculation is applied to the data of experiment 1 (table 1) where $x_0 = 117$, $x_f = 5.7$, $Va = 0.113$ mM and the glucose decomposed is 0.3 mM, a result of 1.14 moles of carbon dioxide per mole of glucose fermented is obtained. This figure almost certainly errs on the low side because of the likelihood that x_f was raised by contamination with $C^{*}O_2$ from the gas phase of the culture vessel. Such contamination might easily have been important in this experiment due to the very low activity of the final carbon dioxide and the high activity of the initial carbon dioxide, some of which may have remained in the gas phase or in the Oxsorbent seal during the fermentation. When x_f is very small in relation to x_0, a relatively slight increase in x_f will cause a disproportionately large decrease in the factor $\log x_0/x_f$ and therefore in V_f.

To obtain a more reliable value for carbon dioxide production two additional experiments were performed in which the ratio x_0/x_f was kept small by limiting the amount of glucose fermented. In this way errors due to C^{14} contamination were made insignificant. From the data of these experi-

ments (table 4) the intermediate carbon dioxide production was calculated to be 2.32 (experiment 2) and 2.19 (experiment 3) moles per mole of glucose fermented.

TABLE 4

THE FERMENTATION OF GLUCOSE IN THE PRESENCE OF C^*O_2

Experiments 2 and 3

	EXPT. 2		EXPT. 3	
COMPOUND	MM/10 ML.	CTS./MIN./MM $\times 10^{-3}$	MM/10 ML.	CTS./MIN./MM $\times 10^{-3}$
Carbon dioxide, initial	(0.191)	28.5 ± 0.5	(0.136)	25.5 ± 0.4
Carbon dioxide, final	0.191	10.0 ± 0.3	0.136	6.9 ± 0.3
Acetic acid, final	0.258	9.5 ± 0.2	0.245	8.6 ± 0.2
Glucose fermented	0.086	....	0.081	....

The above results definitely exclude a mechanism such as that implied in reactions (3) and (4). The agreement with the value of 2 to be expected on the basis of reactions (1) and (2) is very satisfactory in the light of the possibility that reactions other than those postulated may be contributing to the dilution of the C^*O_2. It should be noted in this connection that only about 85% of the glucose is accounted for as acetic acid; the remainder goes into cell material and unidentified non-volatile compounds.

It may be concluded that the "acetic acid fermentation" of glucose by *Cl. thermoaceticum* involves a partial oxidation of the substrate to two moles each of acetic acid and carbon dioxide followed by a reduction and condensation of the carbon dioxide to a third mole of acetic acid. *Cl. thermoaceticum* is the third species of Clostridium that has been shown to use carbon dioxide in this way.

The authors are greatly indebted to Prof. E. O. Lawrence and members of the staff of the Radiation Laboratory and to Dr. T. H. Norris of the Department of Chemistry for facilities used in the preparation of C^{14}.

* Present address: Mallinckrodt Institute of Radiology, Washington University, St. Louis, Mo.

† The authors are indebted to Dr. T. H. Norris of the Dept. of Chemistry of the University of California for the data on self-absorption.

[1] Fontaine, F. E., Peterson, W. H., McCoy, E., Johnson, M. J., and Ritter, G. J., *Jour. Bact.*, **43**, 701–715 (1942).

[2] Barker, H. A., these PROCEEDINGS, **30**, 88–90 (1944).

[3] Wieringa, K. T., *Leeuwenhoek*, **3**, 1–11 (1936).

[4] Wieringa, K. T., *Ibid.*, **6**, 251–262 (1939–1940).

[5] Barker, H. A., Ruben, S., and Beck, J. V., these PROCEEDINGS, **26**, 477–482 (1940).

[6] Ruben, S., and Kamen, M. D., *Phys. Rev.*, **57**, 549 (1940).

[7] Kamen, M. D., unpublished data.

[8] Slade, H. D., Wood, H. G., Nier, A. O., Hemingway, A., and Werkman, C. H., *Jour. Biol. Chem.*, **143**, 133–145 (1942).

THE CRYSTALLINE PHASES OF SOAP

By M. J. Buerger, L. B. Smith, F. V. Ryer and J. E. Spike, Jr.

Mineralogical Laboratory, Massachusetts Institute of Technology, and Lever Bros. Research Laboratory

Communicated June 14, 1945

Introduction.—The view has recently been expressed that soap exists in only four solid phases,[6] known as α, β, δ and ω, and that only three of these, namely β, δ and ω, are encountered in commercial soaps. This simplified scheme does not agree with the results of our own studies of the crystalline phases of soap. We believe that at least part of the error of this view is due to the fact that some of the supposedly simple phases described are actually composites, and that part of it is due to limited knowledge of soap phases themselves. Since we feel that the understanding of soap phases is in a very confused state, we take this opportunity to present information which will tend to clear the existing confusion.

Orthodox Methods of Producing Various Phases.—Thiessen and Stauff[1] were the first to point out that sodium stearate and sodium palmitate occur in more than one form. They obtained a form, which they designated α, by crystallization from an alcohol solution. They showed that if this form of sodium stearate is heated, it undergoes an irreversible transformation at 52°C. to another form which they called β. De Bretteville and McBain[2] produced a third form, which they called γ, to conform with the notation of Thiessen and Stauff, by precipitating sodium stearate from 95% alcohol and drying at 105°C. Immediately after this we[3] showed that the forms known as α and β were actually different hydrates, and that the α form can be transformed into β, γ, or a still different form (not then dignified by a label) by heating it within the temperature brackets[4] 52–105°, 105–117° and above 117°C., under otherwise atmospheric conditions. For sake of a label we now propose the designation σ for the form assumed by heating neutral sodium stearate above 117°C.

Shortly after this, Ferguson, Rosevear and Stillman[5] recognized the existence of only the four phases α, β, ω and δ. They obtained their ω phase by cooling neat soap without agitation. They found this phase to occur in commercial framed soaps, most commercial milled soaps, and to predominate in cocoanut oil soaps. They also found it to be the only stable phase in sodium oleate soaps. Another phase which they believed to be identical with Thiessen and Stauff's β, and which they therefore also called β, was formed in those soaps which would otherwise crystallize to form ω, but which were agitated at certain temperatures where it was found that β was stable. The fourth form, δ, they obtained by several methods, including quenching a hydrous melt, mixing α, β or ω with water at room temperature, or extrusion of hydrous sodium palmitate from an orifice at room temperatures.

New Methods of Producing Different Soap Phases.—Soap makers had realized for some time that soaps differed in properties according to water content, and, more recently, Bodman[7] discovered that they differed in properties according to temperature of treatment as well as water content. We conceived that the differences so manifested in commercial soaps were to be accounted for by different content of crystalline phases with different conditions of preparation. At a very early date we proved that this was indeed the case by finding that the diffraction patterns of the several commercial varieties of soap were different and that the diffraction effects arose from crystalline phases. This immediately suggested that the phase diagrams of soap-water systems governed the phase mixtures found in commercial soaps, and so influenced the properties of soaps. Our subsequent experimental work proved this to be true. We therefore explored a number of soap-water systems for phase distribution. Before entering into this, we wish to discuss the concepts of *descendent phases* and *phase maps*.

Descendent Phases.—At this point we wish to emphasize the distinction between a phase stable in a region where it is developed and the phase or phases which descend from it as a result of changing conditions, such as lowering the temperature in order that the phase can be examined under normal "room temperature" conditions. The former phase properly belongs on a phase diagram. The latter phases, here designated *descendent phases*, are not necessarily the same phase which formed under the original conditions, but are the phases which form from the original one during the change in environment. If no transformations occur during the change, then the original phase and the descendent phase are identical; if a transformation does occur, then they are different. As an example of this relation, McBain's subwaxy phase[4] is the original phase, the σ phase is its descendent.

Phase Maps.—If the phases of a phase diagram are replaced at all points by their descendent phases, we designate the resulting diagram as a *phase map*. In other words, a phase map is a diagram showing the room temperature equivalents of the phases on the phase diagram.

Clearly, a phase map of a soap-water system is a type of representation of the phase diagram of the system. It should reveal possible new room temperature soap phases at the same time giving some idea of their relationships to one another. To study this, we explored the phase maps of the following individual pure soaps with water:

SOAP	COMPOSITION	SYMBOL
Sodium stearate	$NaC_{18}H_{36}O_2$	NaC_{18}
Sodium palmitate	$NaC_{16}H_{32}O_2$	NaC_{16}
Sodium myristate	$NaC_{14}H_{28}O_2$	NaC_{14}
Sodium laurate	$NaC_{12}H_{24}O_2$	NaC_{12}
Sodium caprate	$NaC_{10}H_{20}O_2$	NaC_{10}
Sodium oleate	$NaC_{18}H_{34}O_2$	$NaC_{\overset{=}{18}}$

Equilibrium was attained by the process of severely working the sample. This imposes intense plastic deformation on any solid phases. We have found that this provides the necessary activation energy in mechanical form to induce transformations which may otherwise be inhibited.

New Phases.—In this study we encountered seven distinct crystalline soap phases (in addition to several other phases closely related to these). These seven are distributed among the several soap-water systems according to the scheme in table 1.

TABLE 1

THE DISTRIBUTION OF PHASES IN THE PHASE MAPS OF NEUTRAL SODIUM SOAPS WITH WATER

	NaC_{10}	NaC_{12}	NaC_{14}	NaC_{16}	NaC_{18}	$NaC_{\overline{\overline{18}}}$
κ	x	x	x	..	..	..
ζ	..	x	x	x	x	..
μ	..	..	x	..	..	..
ϵ	..	..	..	x	x	..
δ	..	.	..	x	x	..
α	..	..	..	..	x	..
η	..	..	..	..	..	x

It will be noted that Ferguson, Rosevear and Stillman's[5] ω and β phases are missing from this list. The significance of this is that each of these phases is composite, and we resolve them into the following components:

Ferguson, Rosevear and Stillman's phases		Our phases
ω	$\longrightarrow$	κ and η
β	$\longrightarrow$	ϵ and ζ

Ferguson, Rosevear and Stillman's descriptions of conditions under which their phases occur include conditions under which both components appear. Furthermore, Ferguson's "characterizing diffraction ring" does not distinguish between the components in the case of either of their phases ω or β. That the components are indeed individual phases is attested by the fact that they occur individually in certain fields of certain phase diagrams and together in others. We accordingly reject Ferguson, Rosevear and Stillman's ω and β phases as being each resolvable into at least two phases.

EXPLANATION OF FIGURE 1

X-ray powder photographs of sodium soap phases. These photographs, which are reproduced here to exactly natural size, were made with a camera having the standard diameter of 114.6 mm. (on which 1-mm. film distance corresponds to 1° for the deviation angle 2θ, and to $^1/_2$° for the Bragg angle, θ).

1. α sodium stearate	6. μ sodium myristate
2. δ sodium stearate	7. κ sodium myristate
3. ϵ sodium stearate	8. γ sodium stearate
4. ζ sodium stearate	9. σ sodium stearate
5. β sodium stearate	10. η sodium oleate

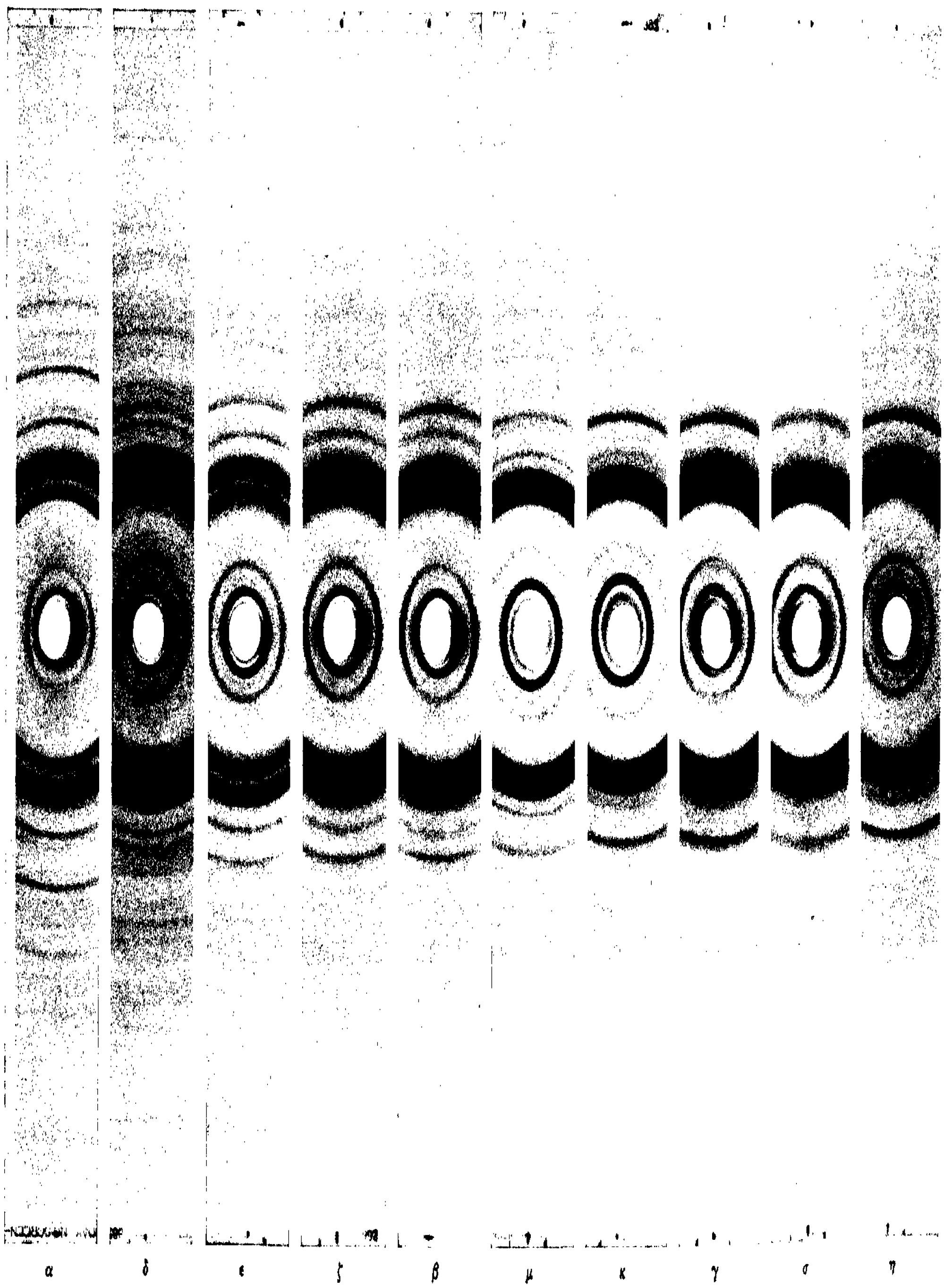


Figure 1

(See opposite page for explanation)

(If the single "characterizing diffraction ring" were the only distinguishing charactertistic of a phase, it would be impossible, with Ferguson's ω data, to distinguish between the four phases κ, η, γ and σ.)

The Rôle of Hydration.—Elsewhere[8] it is shown that all the phases of table 1, which occur in the central areas of phase maps, are hydrates. The amount of hydration is a definite fraction of a mole of water per mole of soap. The presence of definite amounts of water in soap crystals which form in a water environment is a consequence of the requirement that the alkali atom of the soap molecule surrounds itself with its normal coördination sphere of oxygen atoms.[9]

There exist two classes of transformations accompanying the dehydration of soap phases. One class (for example, Thiessen and Stauff's $\alpha \rightarrow \beta$) is accompanied by a complete change of crystal structure. The other class is accompanied by comparatively minor alteration in the structure, and consequently is indicated by a comparatively slight change in the x-ray powder pattern of the original phase. This behavior is true of the dehydration of worked δ, ϵ and ζ.

Since the more and less hydrated modifications of a soap crystal type can be distinguished, even though their powder patterns are rather similar, it is desirable to refer to them in a distinctive manner. We suggest that these forms be distinguished by adding a prime to the symbol of the less hydrated modification. Using this symbolism, we have found that the first step in the dehydration of δ is δ', of ϵ is ϵ' and of ζ is ζ'. Thiessen and Stauff's β is equivalent to ζ'.

We now call attention to certain relations between Ferguson, Rosevear and Stillman's forms and other forms. They originally* believed[5] their β to be identical with Thiessen and Stauff's β. One component of their β is ζ. This component, on dehydration, first loses $^1/_8$ molecule of water and becomes ζ', which is Thiessen and Stauff's β. Even if Ferguson's β were the pure component ζ, it would not be identical with Thiessen and Stauff's original β, because the latter is a product of partial dehydration of the former, and the water is lost in a sharp step, giving no evidence of continuity between the phases. The two phases are also readily distinguished by powder pattern.

Ferguson, Rosevear and Stillman[5] express belief that their ω is identical with de Bretteville and McBain's[2] γ. This can hardly be the case, even if ω were a single phase and not resolvable into κ and η, for both of the latter phases are hydrated,[8] while the former is almost, if not quite, anhydrous.

In order that others may identify the several phases involved in soaps, we present in figure 1 a series of representative x-ray powder photographs of the several phase types encountered in sodium soaps. The patterns of the same phase type but different chain length resemble one another basically because the Fourier transforms of the crystals resemble one another. In

detail, the crystals of similar phase type are nearly identical in projection normal to the direction of the chains in the structure. This direction is undoubtedly nearly normal to the sheet structure of soap crystals. Since the plane of the sheets is customarily designated[9] (001), this means that the crystals are nearly similar when projected along zone [001], whose reflections are *hk0*. Thus the (001)* levels of the reciprocal lattices of crystals of different chain length homologues of the same phase are nearly identical and consequently these phases give identical *hk0* contributions to a powder photograph. The reflections involving the *l* index, however (both 00*l* and *hkl*), differ for patterns of the various chain lengths.

Distribution of Phases on Phase Maps.—In subsequent publications we plan to describe in detail the phase maps of the several systems of pure soaps with water. At this time we merely outline their several characteristics. In figure 2 is shown the phase map of the system NaC_{16}-water, omitting the liquid solution phase which occurs increasingly toward the right of the diagram.

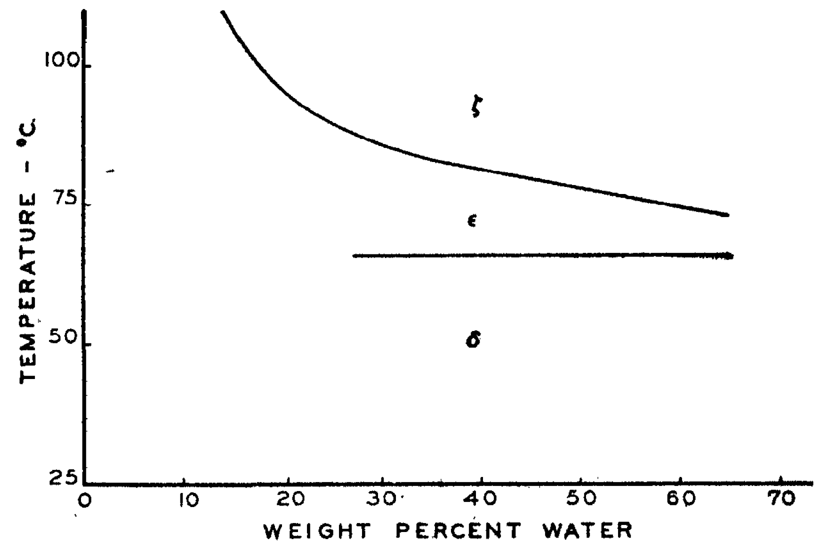


Figure 2

The NaC_{16}-water phase map is one of intermediate complexity, but it illustrates some common features of phase maps of simple soap systems. A major feature of such maps is a set of three regions, the middle region being bounded above by a sloping line and below by an almost horizontal line. These boundaries converge toward higher water contents. The upper line, at least near the middle region, checks well with McBain's[10] T_c, the temperature of the final melting of the crystalline phases. The lower horizontal phase boundary is a new one and evidently represents a phase

transformation in the crystalline state. In the NaC_{16}-water phase map, the δ phase occupies the field below this boundary, the ϵ phase occupies the middle field and the ζ phase occupies the field above the upper boundary. It is visually evident whether a sample has been prepared above or below the ϵ-ζ boundary. Samples worked in the ζ field are brittle and have a compact texture, while those worked in the ϵ field are tougher and have a fibrous appearance with silky luster. The δ-ϵ boundary corresponds well with the temperature of ready solubility, as listed by McBain and Lee.[10]

Phase maps of water with soaps of longer chain length are somewhat more complex. The NaC_{18}-water diagram is similar to the NaC_{16}-water diagram, with the following additional complexities. The δ phase field is a thick, almost horizontal stratum below whose lower boundary occurs another field containing the α phase. Furthermore, the upper field which contains only pure ζ for NaC_{18} is subdivided for NaC_{16}: pure ζ occurs at low moistures, but $\zeta + \delta$ appears at intermediate moistures. Since soap solution is also present, this appears to be a field of three phases, which cannot coexist in the phase diagram of a two-component system. Evidently δ is generated from the phases stable in this region when the sample is cooled for examination.

Phase maps of water with soaps of shorter chain length tend to be simpler. In the NaC_{14}-water phase map, the upper, middle and lower regions contain κ, ζ and μ, respectively. The latter phase occurs only, so far as we know, in this system. (At very low water contents this phase appears to be displaced by κ.) The middle region tapers off at higher moisture contents toward about 70% water. In the NaC_{12}-water map, fewer phases occur. Both upper and lower region are occupied by the κ phase; the middle is occupied by ζ and vanishes at about 75% water. In the NaC_{10}-water system, only the κ phase occurs, yet the middle region is marked out by slightly different detail in the κ powder pattern, as well as by orientation differences.

The phase map of NaOl-water is also very simple, consisting chiefly of the phase η. The only detail in the map corresponds with a minor pattern change below 6% water. We show elsewhere[8] that the samples having water contents greater than this are actually crystals of the hemihydrate, $NaC_{\overline{18}} \cdot {}^1/_2H_2O$. Below 6% water the dehydrated phase η' occurs.

Phase Maps of Commercial Soaps.—Commercial soaps have complex phase maps which differ from soap to soap. This might be expected, since commercial soaps are accidental and convenient mixtures made from fats containing varying spectra of chain lengths, and containing both saturated and unsaturated components.

We are not in agreement with the statement[6] that commercial soaps contain only the three phases β, ω and δ. We find that phase maps of commercial soaps are so complex that we cannot deduce them with any cer-

tainty from powder photographs. In figure 3 we present a phase map of a certain commercial soap. In this map, where we have felt uncertain of identification of the phases of a phase field, we have attempted to indicate the minimum number of pure phase patterns into which we believe the powder photographs can be resolved. We present this diagram with some hesitation, since we are certain we would revise it in light of subsequent knowledge, yet we feel that it should be presented in view of the published statements as to the simplicity of the phase aggregates in commercial soaps.

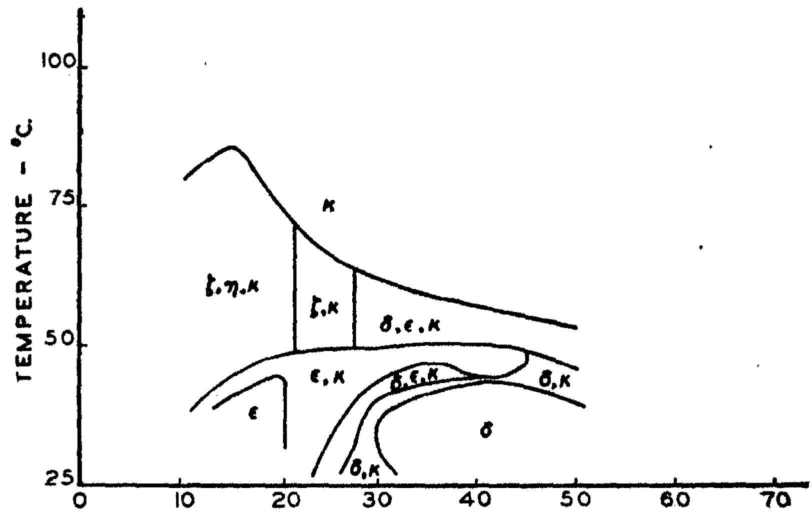


Figure 3

* Subsequently Ferguson[6] expressed the view that his β differed from Thiessen and Stauff's β by being indefinitely hydrated.

[1] Thiessen, Peter A., and Stauff, Joachim, "Feinbau und Umwandlungen kristallisierter Alkalisalze langkettiger Fettsäuren," *Zeit. Physikal. Chem.* (*A*), **176**, 397–429 (1936).

[2] Bretteville, Alexander de, Jr., and McBain, J. W., "X-ray Evidence for a Third Polymorphic Form of Sodium Stearate," *Science*, **96**, 470–471 (1942).

[3] Buerger, M. J., Smith, L. B., de Bretteville, Jr., A., and Ryer, F. V., "The Lower Hydrates of Soap," *Proc. Nat. Acad. Sci.*, **28**, 526–529 (1942).

[4] McBain, James W., Vold, Robert D., and Frick, Mary, "A Phase Rule Study of the System Sodium Stearate-Water," *Jour. Phys. Chem.*, **44**, 1013–1024 (1940).

[5] Ferguson, R. H., Rosevear, F. B., and Stillman, R. C., "Solid Soap Phases," *Ind. Eng. Chem.*, **35**, 1005–1012 (1943).

[6] Ferguson, R. H., "The Four Known Crystalline Forms of Soap," *Oil and Soap*, **21**, 6–9 (1944).

[7] Bodman, John W., U. S. Patent No. 2,215,539, Sept. 24, 1940.

[8] Gardiner, K. W., Buerger, M. J., and Smith, L. B., "The Hydrate Nature of Soap," *Jour. Phys. Chem.* (in press).

[9] Buerger, M. J., "Soap Crystals," *American Mineralogist* (in press).

[10] McBain, James W., and Lee, W. W., "Vapor Pressure Data and Phase Diagrams for Some Concentrated Soap-Water System Above Room Temperature," *Oil and Soap*, 20, 17–25 (1943).

WILLARD GIBBS ON SOARING FLIGHT

BY EDWIN B. WILSON

HARVARD SCHOOL OF PUBLIC HEALTH

Communicated July 12, 1945

In these days when so much depends on aerial supremacy it may interest some to see a hitherto unpublished letter of Gibbs to Langley dating from the very early infancy of aeronautical research in this country. The letter was furnished to me by Professor Ralph G. VanName, nephew of Gibbs.

On April 13, 1894, Langley wrote Gibbs asking for help in understanding soaring flight. He gave the usual Newton-type of formula for the wind pressure upon the aerofoil, proportional to the area, the square of the velocity and a function of the angle of attack which could be taken as proportional to that angle when it was small. Six weeks later Gibbs replied as follows:

New Haven, May 30/94

My dear Professor Langley

I do not know whether the following results—certainly very meager after so long a delay—will be at all to your purpose, but the discussion of a simple case may throw some light on the general question.

Let the velocity of the wind be

$$V = A + a \sin nt. \tag{1}$$

Let the horizontal velocity of the aeroplane (measured *against* the wind) be regulated by varying the inclination so as to be expressed by a similar function, say,

$$v = B + b \cos nt. \tag{2}$$

The *relative* horizontal velocity will be

$$w = C + a \sin nt + b \cos nt, \tag{3}$$

where $C = A + B$. If α is the tangent of the elevation of the edge (supposed small),

$$\alpha = -\frac{dv}{gdt} - \frac{Ew^2}{g}, \tag{4}$$

where Ew^2 is the resistance of the air to the edge of the plane divided by its mass; and if z is the vertical height,

$$\frac{dz}{dt} = \alpha w - \frac{g}{Pw} = -\frac{w}{g}\frac{dv}{dt} - \frac{Ew^3}{g} - \frac{g}{Pw}, \tag{5}$$

where g/Pw is the rate at which the aeroplane would settle down when $\alpha = 0$. We shall treat P as roughly a constant. It is here assumed not only that α is small, but also that w is large, and d^2z/dt^2 is small compared with g, so that the variations of vertical momentum may be neglected in (4). Since by (2)

$$\frac{dv}{dt} = -nb \sin nt, \tag{6}$$

(5), (6) and (3) give

$$\frac{dz}{dt} = \frac{bn}{g} \sin nt\ (C + a \sin nt + b \cos nt) - \frac{Ew^3}{g} - \frac{g}{Pw}, \tag{7}$$

$$\int_{t=0}^{t=2\pi/n} dz = \frac{\pi ab}{g} - \frac{2\pi}{n}\left[\frac{Ew^3}{g} + \frac{g}{Pw}\right] \text{average value.} \tag{8}$$

Continued soaring is possible, if

$$\frac{nab}{2g} > \left(\frac{Ew^3}{g} + \frac{g}{Pw}\right) \text{ave.} \tag{9}$$

Since C is the average value of w, we shall err on the side of safety if we write for (9),

$$\frac{nab}{6g} > \frac{EC^3}{g} + \frac{g}{PC}, \tag{10}$$

provided that C is three times as great as a and as b, and *provided* also that α and $\frac{1}{g}\frac{d^2z}{dt^2}$ are sufficiently small to make our formulae reasonably accurate.

To make α small, nb must be small compared with g. To make d^2z/dt^2 small compared with g, we must have n^2bC small compared with g^2. To fix our ideas, we may make $b = C/3$. Then (10) becomes

$$\frac{na}{18g} > \frac{EC^2}{g} + \frac{g}{PC^2}. \tag{11}$$

The left-hand member of this condition is given by the wind. If E is small enough and P large enough, it is easy to satisfy (11) by values of C which

are small compared to g/n, which will make soaring possible, and if such values of C or any of them are greater than A (the average velocity of the wind), continued soaring *against* the wind is possible.

The letter is taken from the handwritten copy which Gibbs kept and may not be identical with that which he sent; it closes without the customary "Yours truly" or equivalent and without signature, though undoubtedly both appeared on the letter sent to Langley.

Under date of June 9, 1894, Langley wrote: "I beg to return my very best thanks for your highly valued communication of the 31st ultimo. Not being certain when I should hear from you, I had asked the independent assistance of one or two other gentlemen, whose analytical skill I thought might be more trusted than my own.

"I am chiefly surprised at the entirely different ways in which the problem can be looked at and treated, and yours is certainly original and distinctive. It has arrived too late, I regret, to be used in a communication which is just going to France, but I shall probably make use of it later, with due public acknowledgment of your valued aid and great kindness."

About six months later (January 25, 1895), Langley wrote again:

"You will be interested in the discussion of a mathematical problem in aerodynamics, prepared at my request by Mr. de Saussure independently of the discussion that you were good enough to send me, which I have preserved, and may ask your permission to make use of it at some future time. Mr. de Saussure's article appeared as a supplement to the French translation of my paper on The Internal Work of the Wind in the *Revue de l'Aeronautique.*

"I have sent to you, under a separate cover, a copy of the paper in question, and, should your leisure permit, I shall be glad to hear any criticism upon it that you may make.

"Renewing my thanks for your assistance, I am,"

So far as I have been able to ascertain Langley never took occasion to use Gibbs's material, nor Gibbs to accept the invitation to comment on de Saussure's article—there is certainly no evidence of either in the meager papers Gibbs left at his death.

EFFECTS OF EXPOSURE TO ULTRA-VIOLET LIGHT ON VISUAL THRESHOLDS*

By Ernst Wolf

Biological Laboratories, Harvard University

Communicated July 5, 1945

Effects of ultra-violet light upon the eye are well known in the form of "snow blindness" and in many other pathological appearances found particularly among industrial workers who are exposed to light sources emitting great quantities of ultra-violet.[1, 2, 3] Since the wave band of light is continuous from abiotic rays to the visible part of the spectrum, it is of interest to know the limits of detrimental effectiveness of the short wave-lengths.

The transmission of the ocular media has been studied by scores of investigators and the observations vary considerably. We may assume, however, that some human and other vertebrate eyes transmit wave-lengths as low as 310 mμ. Recent visual tests in the ultra-violet[4] indicate that light of 302 mμ is perceived by very young subjects (6–10 years). With progressing age the low limit moves to longer wave-lengths, middle-aged persons rising to 310–320 mμ and at higher age to 360 mμ and higher. In laboratory tests a decrease in visual acuity after exposure to ultra-violet has been found in monkeys and man.[5] No conclusions concerning wave-length limits causing these effects can be drawn from these data. For industry, suggestions regarding safety measures have been made in accordance with the assumption that only wave-lengths shorter than 305 mμ would be detrimental to the visual mechanism and that crown or flint glass of sufficient thickness would provide adequate protection.[6] A systematic study of the effects of different wave bands below 400 mμ on the vision of some suitable animal material which would permit comparison with similar effects upon the human eye seems therefore desirable.

Responses to flicker have been studied in a great variety of vertebrates and flicker functions established.[7] In animals the end-point of flicker recognition at a given intensity of light is determined by the appearance of a head nystagmus due to the perception of a cylindrical system of alternate translucent and opaque stripes rotating around the animal.[8] While the test method differs from that customarily used for man, the flicker function obtained in this fashion is identical with such functions determined by other methods.[9, 10] The behavior of flicker thresholds in relation to light intensity permits therefore in any suitable organism an investigation of visual function under varied experimental conditions, for instance after exposure to ultra-violet light. If by the ultra-violet radiation any functional changes are induced, it is possible to demonstrate (a) any deviation

of a response threshold from its normal level at any given point along the flicker function, (*b*) its magnitude of deviation in relation to exposure time and spectral region employed and (*c*) the course of return of the threshold to its normal level with time.

For test, an animal was chosen which would fulfil the following conditions: (*a*) availability in quantity at all seasons, (*b*) genetic uniformity, (*c*) precision and reproducibility of a threshold response and (*d*) an organism that would without aid keep its eyes open during exposure to ultra-violet, so that a full effect of irradiation is assured. For the purpose baby chicks, 3–15 days old, of white Leghorn stock, proved particularly suitable.

The eye of the chick transmits at least as far as 315 mμ. The visual purple of the chick and its absorption spectrum are identical with those of man.[11] The chick's eye possesses, however, embedded in the retina considerable numbers of colored oil droplets which may act as additional light filters and which are absent in the human eye.[12] But since such a screen would rather reduce than enhance the action of ultra-violet, it is justifiable to assume that in other eyes the ultra-violet effect might be more pronounced.

For normal chicks a flicker function has been established[13] which bears the typical characteristics of such curves for other visually duplex vertebrates.[14] The critical flicker frequency is a function of the intensity of illumination, increasing in a double S-shaped form with light intensity, according to the participation of rods and cones in the visual process (Fig. 1). For testing F = 30/sec. with a light/dark ratio of 1:1 in each flicker cycle has been chosen, and the change of threshold value studied after the chick's exposure to ultra-violet. The testing point lies in the midregion of the curve and at an intensity level where responses are precise and easily recognizable by the observer.

While the chicks (4 to 6 at one time) are confined to a cylindrical wire mesh cage, but free to move, with access to food and water, they are exposed to the intense radiation of three 250-watt mercury vapor lamps of the GE type H-5 which are mounted at eye level equally spaced outside the cage and surrounded by a reflecting surface. Lamp housings permit the insertion of filters for selection of specific spectral regions. For exposure, a standard period of 60 minutes was chosen.

Each individual chick is then transferred to a glass cage to be tested at F = 30/sec. for the onset of the nystagmus response in dependence on light intensity (white light), and the tests are repeated at intervals until the threshold returns to its normal level. The first reading is taken 60 minutes after cessation of exposure to allow for complete dark adaptation. Dark adaptation should be complete after 40 to 45 minutes;[15] any deviation of the threshold from its normal level should therefore be due to the exposure to the ultra-violet and not to incomplete adaptation. When

exposing for control some chicks, not previously exposed to ultra-violet, to *white* light (Tungsten filament lamps), equal in intensity to the three mercury vapor lamps, and after proper dark adaptation, no effects on the threshold can be noticed. After ultra-violet treatment the threshold is, however, much higher, depending upon the extent of the spectral region employed.

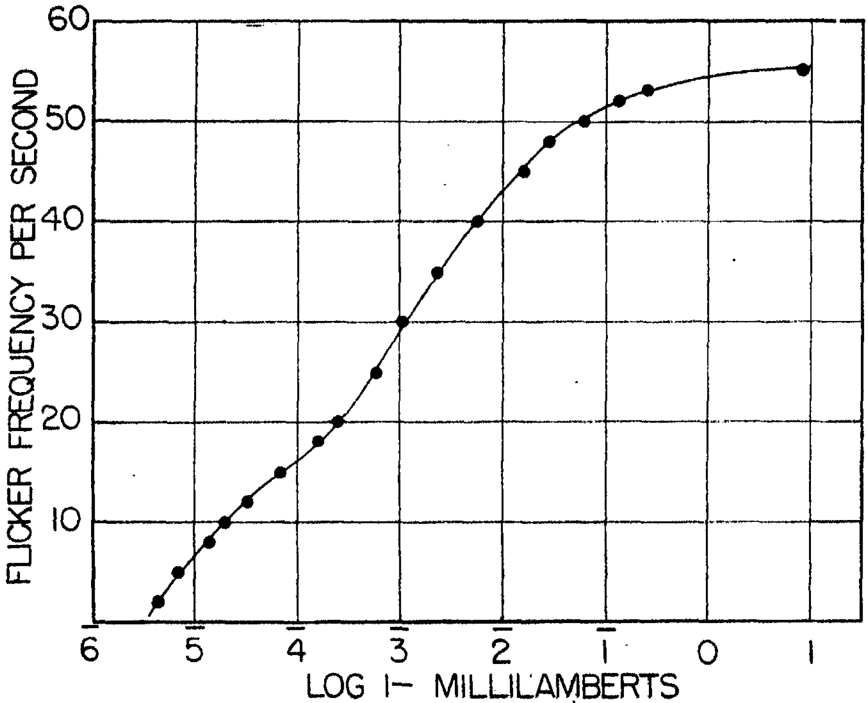


FIGURE 1

Relations between flicker frequency (F) and mean critical intensity of light (I) for response to flicker in the chick.

Chicks receiving the full impact of the mercury vapor lamps from which the surrounding glass jackets have been removed may show conjunctivitis, keratitis and hyperemia of corneal capillaries, while the fundus looks normal. These pathological changes are, however, absent as soon as the quartz mercury vapor tubes are shielded (as they usually are) by their envelopes of Corning glass No. 774. A study of the behavior of the flicker recognition end-point reveals, however, that it is greatly altered and takes many hours to return to its normal level. It is, therefore, evident that the exterior appearance of the eye does not suffice to decide whether the visual mechanism has been affected by the ultra-violet.

The spectral range of particular interest lies between the lowest limit of transmission of ultra-violet by the eye (>300 mμ) and the beginning of the visible range (<400 mμ). By the choice of suitable filters this region is subdivided to ascertain the specific ultra-violet bands which affect the visual threshold. The filters according to their order of transmission are: Corning 774, AO crown glass 1045, ordinary plate glass, AO Cruxite 1794, Corning 3850, AO Calobar C 1827, AO 2614, Polaroid XY 91 VO and Corning 3389. Their transmission curves, determined with a Beckman spectrophotometer, are given in figure 2.

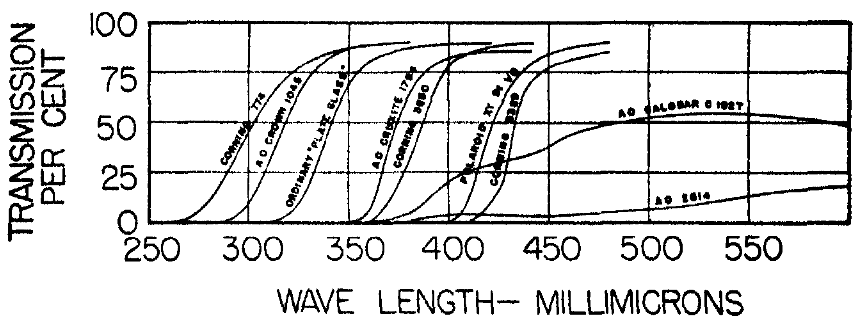


FIGURE 2

Spectrophotometric transmissions of filter glasses used for elimination of the ultra-violet part of the spectrum.

With the mercury vapor as light source and shielded by quartz only, the first threshold recorded one hour after cessation of exposure is about 45 times higher than the normal. Repeated tests show a gradual lowering of the threshold with time, until the normal level is reached not before 72 hours (3 days). In this case pathological changes are clearly visible and some damaging effects were to be expected. With Corning 774 as a filter, a smaller initial threshold change is obtained. This filter begins to transmit at 260 mμ and thus excludes the strong 254 mμ mercury line. In this case the amount of light necessary for threshold recognition is 5 times the normal, and the approach to the normal sensitivity level is complete after about 30 hours. By gradually reducing the extent of the ultra-violet band from 290 mμ to 360 mμ through the use of crown, plate glass, Cruxite and Corning 3850, the initial threshold change and the subsequent recovery are found to be direct functions of the gradual elimination of the short waves. For crown glass the initial threshold is 4 times, for plate glass 3 times, for Cruxite 1.5 times and for Corning 3850 1.3 times higher than normal; and for crown, plate, Cruxite and Corning 3850 the final level is reached after 24, 18, 9 and 6 hours, respectively.

With filters which do not transmit any appreciable amount of light at 365 mμ and below, the picture changes abruptly. The first threshold

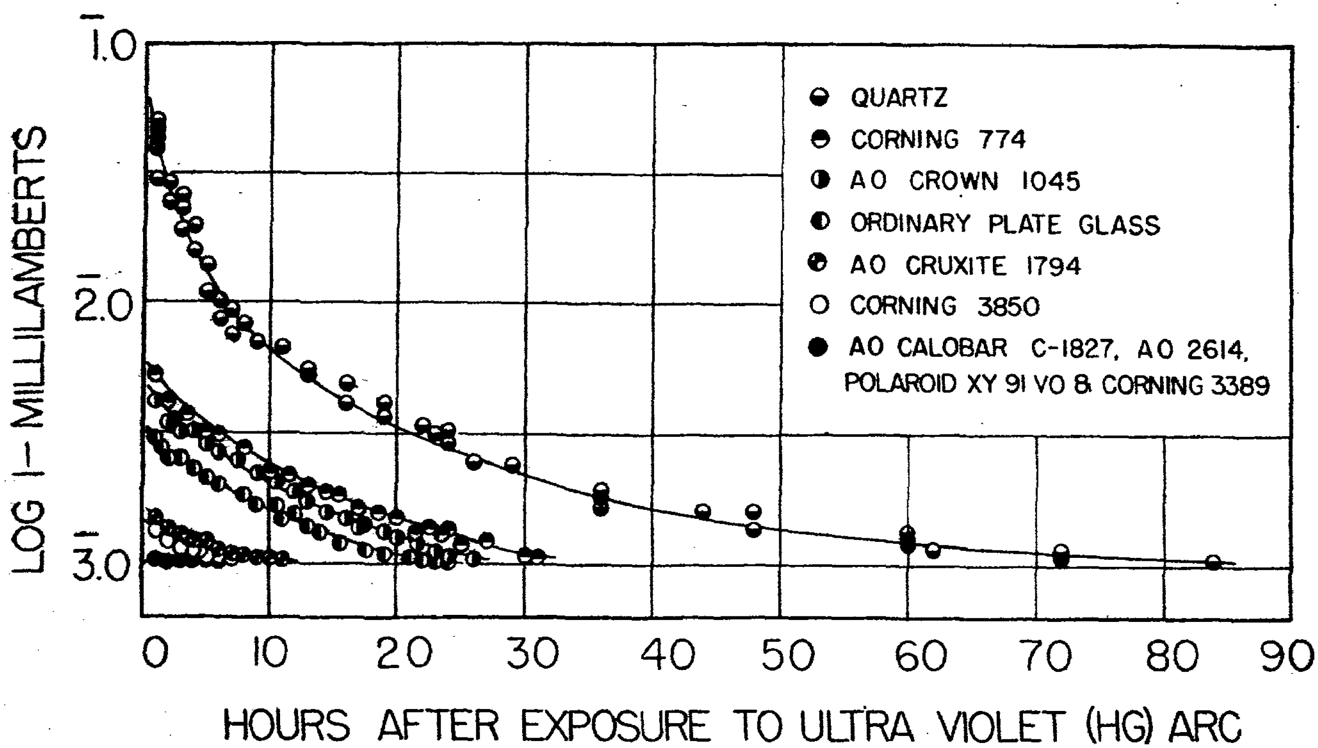


FIGURE 3

Determinations of threshold intensities of light necessary for recognition of flicker at F = 30/sec. after exposure to ultra-violet light of different spectral regions. One hour after completion of exposure the thresholds are highest and decrease with time until a final level is reached. The initial deviation from the normal and the time for recovery are functions of the extent of the ultra-violet spectrum.

determination one hour after completion of exposure is at the normal level and does not change in repeated tests. AO 2614 begins to transmit at 380 mμ, Polaroid XY 91 VO and Corning 3389 at 400 mμ and above. Calobar transmits about 0.5% at 360 mμ and 3.1% at 370 mμ. The mercury spectrum has very strong lines at about 365 mμ which probably represent the limit of injurious radiation. For Calobar the relative energy transmitted in this region is only one-twentieth of that transmitted by Corning 3850; consequently a small ultra-violet effect is apparent with the Corning filter and not with Calobar.[16] The data are shown in detail in figure 3. Each point represents the mean of 4 to 6 individual threshold determinations. The deviation from the mean is in no case greater than $^1/_{10}$, and in most cases not greater than $^1/_{20}$ of a log. unit.

Summary.—If, based upon the similarity of the transmission of the ocular media, conclusions can be drawn from the findings on the eye of the chick, it is apparent that under absence of exterior pathological conditiöns, recognizable by ophthalmoscopic inspection, the visual mechanism is impaired by ultra-violet light between 300 and 365 mμ. Protective means for the eye should, therefore, be filters which absorb the ultra-violet up to 365 mμ totally or sufficiently to prevent any injurious effects.

* This research was supported by a grant of the American Optical Company.

[1] Laurens, H., *The Physiological Effects of Radiant Energy*, Chemical Catalog Co., New York, 1933.

[2] Blum, H. F., *Photodynamic Action and Diseases Caused by Light*, Reinhold, New York, 1941.

[3] Rieke, F. E., *Jour. Am. Med. Assoc.*, **122**, 734 (1943).

[4] Friedrich, W., and Schreiber, H., *Arch. ges. Physiol.*, **246**, 621, 790 (1943).

[5] Verhoeff, F. H., and Bell, L., "The Pathological Effects of Radiant Energy on the Eye," *Proc. Am. Acad. Sci.*, **51**, No. 13 (1916).

[6] Kinsey, V. E., Cogan, D. G., and Drinker, P., *Jour. Am. Med. Assoc.*, **123**, 403 (1943).

[7] Crozier, W. J., Wolf, E., and Zerrahn-Wolf, G., *Jour. Gen. Physiol.*, **20**, 211 (1936); **21**, 17 (1937). Crozier, W. J., and Wolf, E., *Ibid.*, **21**, 313 (1938); **22**, 311, 436, 555 (1939); **23**, 229, 667 (1940); **24**, 317 (1940), 585 (1941); **25**, 331 (1942); **27**, 315 (1944).

[8] Crozier, W. J., and Wolf, E., *Ibid.*, **23**, 531 (1940).

[9] Crozier, W. J., Wolf, E., and Zerrahn-Wolf, G., *Ibid.*, **21**, 203 (1937).

[10] Crozier, W. J., and Wolf, E., *Ibid.*, **27**, 287 (1944).

[11] Wald, G., and Zussman, H., *Jour. Biol. Chem.*, **122**, 449 (1938).

[12] Franz, V., *Handbuch. d. vergl. Anat. d. Wirbeltiere*, Berlin, 1934. Krause, R., *Mikroskopische Anatomie der Wirbeltiere*, Vol. 2, Berlin, 1922. Walls, G. L., *The Vertebrate Eye*, Cranbrook Inst., Bloomfield Hills, Michigan, 1942. Wood, C. A., *The Fundus Oculi of Birds*, Lakeside Press, Chicago, 1917.

[13] Data to be published in detail elsewhere.

[14] Crozier, W. J., Wolf, E., and Zerrahn-Wolf, G., these PROCEEDINGS, **24**, 125, 438 (1938).

[15] Hecht, S., "The Nature of the Photoreceptor Process," Handbook of Gen. Exp. Psychology, Clark University Press, Worcester, 1934.

[16] Filter glasses transmitting ultra-violet and excluding the visible part of the spectrum produce ultra-violet effects similar to those presented above. The details will be discussed in a later publication.

ON PLATEAU'S PROBLEM WITH FREE BOUNDARIES

By R. Courant

New York University

Communicated July 5, 1945

Plateau's problem with free boundaries concerns minimal surfaces M of least area with parts of the boundary prescribed as Jordan arcs while other "free" parts of the boundary are merely restricted to prescribed boundary surfaces S. The existence of such minimal surfaces has been established,[1] but the question remains open whether specific statements concerning the "*trace*" of the minimal surface M on the prescribed boundary surfaces S can be made, "trace" meaning the set of boundary points of M on S. Examples show that this trace need not be a continuous curve;[2] therefore, to secure a "reasonable" trace of M, conditions have to be imposed on S.

To fix the ideas we consider the case of a doubly connected minimal surface M of least area with its free boundary on a closed surface S and its fixed boundary on a closed Jordan curve Γ outside S; M, lying in the $\mathbf{x}$-space, may be parametrically represented by an harmonic vector $\mathbf{x}(u, v)$ with components x, y, z where u, v ranges over an annular ring $\bar{M}$ in the u, v plane between the unit circle $\bar{\rho}$ and a concentric circle $\bar{\Gamma}$, so that $\bar{\Gamma}$ is mapped onto Γ and that $\bar{\rho}$ corresponds to the trace ρ of M on S.

It is the purpose of the present note to show:[3]

(*a*) *The trace is a continuous curve if the prescribed boundary surface S is convex.*

(*b*) *The trace is a rectifiable curve if S is convex and if there exists a cone of supporting planes of S bounding together with S a portion of the space which contains* Γ.[4]

We first prove the theorem (*a*): By a basic elementary lemma[5] one can draw in $\bar{M}$ circular arcs $\bar{\gamma}$ of arbitrarily small radii with any point $\bar{p}$ on $\bar{\rho}$ as center, joining two points on $\bar{\rho}$, such that the images γ in the $\mathbf{x}$-space are arbitrarily short (and analytic, except possibly at the end-points). The arc $\bar{\gamma}$ as well as its image γ on M is called a *bridge*.

The small pieces of the minimal surface M corresponding to the small biangular part of $\bar{M}$ between $\bar{\gamma}$ and $\bar{\rho}$ is denoted by $M(\gamma)$. Then we prove by indirect reasoning: For sufficiently small ϵ there exists a $\delta(\epsilon)$ so that, for all points $\bar{p}$ on $\bar{\rho}$, the piece $M(\gamma)$ is confined in a sphere of radius ϵ provided that the bridge γ is shorter than $\delta(\epsilon)$. If this statement were not true there would exist a sequence of bridges $\bar{\gamma}$ and γ for which the lengths tend to zero and for which the diameter of $M(\gamma)$ remains above a positive bound c. We may assume that the bridges γ converge to a point O on S. About O we draw a sphere Σ of a small radius r which encloses the small bridge. Then we show—and this is sufficient for the proof of our

statement (a)—that $M(\gamma)$ lies entirely in Σ. To this end we exploit the assumed minimum area property of M by performing a geometrical construction which would provide an admissible surface with area less than that of M if we assume that $M(\gamma)$ has points outside of Σ.

Making this assumption we replace $M(\gamma)$ by a surface $M_0(\gamma)$ which consists of the part of $M(\gamma)$ inside Σ and in addition of that part of Σ formed by points onto which points of $M(\gamma)$ outside of Σ and visible from O are projected from O. By normal projection onto a sphere the area of a surface outside is diminished;[6] and because of the convexity of S the surface $M_0(\gamma)$ joins γ with S; hence the surface $M - M(\gamma) + M_0(\gamma)$ would be an admissible surface in the original minimum problem with area less than that of M, contrary to our assumption. Thus theorem (a) is proved.

The proof of theorem (b) requires a more elaborate reasoning. It proceeds in the following steps:

1. The given surface S is approximated by a sequence of convex polyhedra K for which again cones of supporting planes exist as specified above. Without restricting the generality of the result we assume that at every vertex of K three faces converge and form three obtuse angles.

2. With such a polyhedron K instead of S as the free boundary the minimal surface M is shown to have on K a trace ρ which consists of a finite number of analytic arcs.

3. The assumption of theorem (b) permits the conclusion—and this is the key to the proof—that for a sequence of polyhedra K approaching S the length L of the piecewise analytic trace ρ remains uniformly bounded.

4. Now one may choose a subsequence of polyhedra K approaching S such that the boundaries $\rho = \rho_K$ converge to a rectifiable curve ρ on S. According to the established theory, the corresponding minimal surfaces converge to a minimal surface which solves the problem for the free boundary surface S and has as trace on S the rectifiable curve ρ.

Incidentally, by Fatou's theorem, the *tangent plane* on M in a point P has a *limiting position* for P approaching almost every point on ρ and this position is *normal* to S.

As to the details, the first step, being of a quite elementary nature, is not amplified here. Let us for the moment assume the statement (2), then (3) can be proved as follows: First: The—piecewise analytic—trace ρ of M on K (and the whole surface M) lies together with Γ inside of the cone formed by the supporting planes. For, otherwise the minimal surface M of least area connecting Γ with K would protrude across one of the supporting planes, say the plane Π. Since K is convex and since M rests on K this situation would imply the existence of a supporting plane to M, parallel to Π, which is incompatible with the fact that M has negative curvature.

If θ and r denote polar coordinates in the annular ring M then $\mathbf{x}(u, v)$ remains analytic on $\bar{\rho}$, i.e., along $r = 1$, except for a finite number of points. As known from the established theory M is orthogonal to K at points of ρ on a face; likewise along a portion of ρ coinciding with an edge of K the normals to M lie in a supporting plane to K through this edge.[7] If $\bar{\rho}$ is monotonically described the image point on ρ will describe ρ monotonically, as follows from the established theory.

Omitting the finite number of points on $\bar{\rho}$ in which the analyticity of $\mathbf{x}(u, v)$ is interrupted, i.e., which correspond to points on ρ in which an edge or vertex is reached from a face, we now recognize $\mathbf{x}_r$ along ρ as a vector normal to a face or to a supporting plane of K. Taking the vertex of our cone of supporting planes as origin in the $\mathbf{x}$-space, the vectors $\mathbf{x}$ and $-\mathbf{x}_r$ form an acute angle not exceeding $\pi/2 - \alpha$, α being a positive quantity. Hence

$$-\mathbf{x}\mathbf{x}_r > |\mathbf{x}|\ |\mathbf{x}_r| \frac{1}{\sin\alpha}.$$

Since along ρ the distance $|\mathbf{x}|$ is bounded from below we have along ρ an inequality of the form

$$|\mathbf{x}_r| < -A\ \mathbf{x}\mathbf{x}_r \qquad (1)$$

with a value of A which can be taken as the same for all minimal surfaces M belonging to polyhedra K sufficiently near to S in our approximation.

As a sequence of polyhedra K tends to S the corresponding annular rings $\bar{M}$ converge to a limit ring. $\bar{M}'$ may denote the ring between $\bar{\rho}$ and the middle circle $\bar{\rho}'$ of the ring; then $\mathbf{x}$ is analytic on $\bar{\rho}'$ and $|\mathbf{x}|$ and $|\mathbf{x}_r|$ are bounded along $\bar{\rho}'$. Since the area of $M = M_K$ converges to the area of $M = M_S$, the established theory of minimal surfaces implies that

$$D(\mathbf{x}) = \int\int_{\bar{M}} (\mathbf{x}_u^2 + \mathbf{x}_v^2)\, dudv < A_1$$

where also A_1 may be chosen as the same bound for all polyhedra K under consideration. It follows by a well-known reasoning that not only $|\mathbf{x}|$ but also $|\mathbf{x}_r|$ on $\bar{\rho}'$ is bounded by the same constant A_2 for all polyhedra considered.

Using Green's formula and $\Delta\mathbf{x} = 0$ we have for the half ring

$$A_1 > \int\int_{\bar{M}} (\mathbf{x}_u^2 + \mathbf{x}_v^2)\, dudv = -\int_{r=1} \mathbf{x}\mathbf{x}_r\, d\theta + r' \int_{r=r'} \mathbf{x}\mathbf{x}_r\, d\theta$$

where r' is the radius of ρ'. Hence

$$-\int_{r=1} \mathbf{x}\mathbf{x}_r\, d\theta < A_1 + r' \int_{r=r'} |\mathbf{x}|\ |\mathbf{x}_r|\, d\theta.$$

By (1)

$$\int_{r=1} |\mathbf{x}_r|\, d\theta < AA_1 + Ar' \int_{r=r'} |\mathbf{x}|\ |\mathbf{x}_r|\, d\theta$$

or because of the boundedness of $|\mathbf{x}|$ and $|\mathbf{x}_r|$ on $\bar{p}$,

$$\int_{r=1} |\mathbf{x}_r|\, d\theta < A_2 \tag{2}$$

where A_2 is a constant. Now, $\mathbf{x}$ representing a minimal surface, we have on $r = 1$ the relation $|\mathbf{x}_r| = |\mathbf{x}_\theta|$. Hence

$$L = \int_{r=1} |\mathbf{x}_\theta|\, d\theta < A_2 \tag{3}$$

where L is the length of ρ and A_2 is a constant. The statement (3) and consequently the conclusion (4) is thus proved.

It remains to establish beyond (a) the assertion (2) for our polyhedra K: the circle $\bar{p}$ is divided into a finite number of arcs such that each arc is mapped by $\mathbf{x}(u, v)$ either on a segment of an edge, or onto a curve on a face (that the mapping is analytic on these arcs follows from the principle of reflection).

To complete the proof of statement (2) a further chain of arguments is needed.[8]

First some preliminary remarks: In an arbitrary neighborhood of a bridge γ of diameter less than ϵ there are bridges of diameter less than ϵ whose end-points are analytic; for, points on $\bar{p}$ where the (continuous) vector $\mathbf{x}$ is not analytic are nowhere dense and the basic lemma can be applied to bridges γ with end-points in intervals of analyticity.

Second: Lines $x' = ax + by + cz + d = 0$ with constant a, b, c, d are equipotential lines in $\bar{M}$; curves $x' = 0$ in $\bar{M}$ which separate domains $x' > 0$ from $x' < 0$ are unambiguously defined.

Third: A bridge $\bar{\gamma}$ analytic including its end-points can have only a finite number of intersections with curves $x' = 0$ unless $\bar{\gamma}$ itself is a curve $x' = 0$ and hence γ a plane curve.

Now we introduce the concept of a plane T *transversal*[9] to K at a point p: A plane T through p on K with trace τ on K is called transversal if for each bridge[10] β sufficiently near to p and lying in T the plane part of T between β and τ is the surface of least area bounded by β and K. Obviously for p on a face A each plane T through p perpendicular to A is a transversal plane. Furthermore, for p on an edge α each plane through α normal to a supporting plane through α is transversal; finally for a vertex p there exists at least one edge α such that the plane T through α and perpendicular to the opposite face A of the trihedral angle in p is transversal. (This latter statement is a non-trivial consequence of the assumption that all the angles at the edges of K are obtuse.)

To refute the assumption that statement (2) is wrong the transversal planes now are used in the following manner: Suppose first that the trace ρ (continuous by theorem (a)) intersects an edge α infinitely often in the neighborhood of a point p of α, p not a vertex. Let $x' = 0$ be a transversal plane T through p, let $\bar{\gamma}$ and γ be a small bridge (original and image) near

$\bar{p}$ or p; consider a plane curve $x' = 0$ separating a zone $x' > 0$ from a zone $x' < 0$ on $M(\gamma)$ and the corresponding equipotential line in $\bar{M}$. The latter, starting from $\bar{p}$ must end on $\bar{\gamma}$; for otherwise it would form a small bridge $\bar{\beta}$ and consequently $M(\beta)$ would be a part of the plane T, and M would be altogether plane, which can be excluded from the outset. Since infinitely many successive intervals $x' > 0$ and $x' < 0$ on $\bar{p}$ near $\bar{p}$ are assumed we would have infinitely many points on $\bar{\gamma}$ with $x' = 0$, hence $\bar{\gamma}$ would be a line $x' = 0$ which can be excluded by proper choice of $\bar{\gamma}$ unless M is plane.

In a similar way we can rule out the possibility that a point p on α is a point of accumulation of points where ρ reaches α from a face B without crossing. In this case we use transversal planes $T(d)$ normal to B at the distance d from α and then let d tend to zero whereupon a similar contradiction as above is obtained.

If $\bar{p}$ on $\bar{p}$ is mapped on a vertex p and if infinitely many points on $\bar{p}$ near $\bar{p}$ are mapped on p or on points on edges through p while infinitely many other intermediate points are mapped on points in faces, a contradiction is obtained by using the transversal plane $x' = 0$ through p in a way not essentially different from that sketched above.

Combined, these results establish the assertion (2) and thus complete the proof of our theorem.

[1] See, e.g., Courant, "The Existence of Minimal Surfaces under Prescribed Boundary Conditions," *Acta Math.*, 72, 51–97 (1940), in particular, p. 81 ff.

[2] See loc. cit., p. 95.

[3] The brief presentation here will be amplified in a book on *Dirichlet's Principle, Conformal Mapping and Plateau's Problem*.

[4] Naturally, in all the various cases of Plateau's problem with free boundaries similar theorems and proofs apply; moreover, the methods used here indicate the possibility of attacking more general variational problems with free boundaries in more than one dimension.

The author wishes to acknowledge that discussions and correspondence with Dr. C. L. Siegel of the Institute for Advanced Study have played a decisive part in carrying out the present investigation.

[5] See loc. cit., p. 70.

[6] Quite generally any piece of a surface, if projected along the outward normal onto a non-negatively curved surface, will be reduced in area. This remark could be used to generalize the theorem (*b*) by substituting for the cone other non-negatively curved surfaces.

[7] The proof follows from a simple geometrical construction: If M ends in a piece α of an edge and forms an acute angle with an adjacent face A, one could substitute for M an admissible surface of smaller area by replacing a part of M adjacent to α and otherwise bounded by an analytic arc λ joining the end points of α and drawn sufficiently near to α by the cylindrical surface projecting λ onto A.

[8] Details will be given in the forthcoming book.

[9] The concept and its application seem to be capable of generalizations to boundary surfaces S not plane, and to other variational problems.

[10] The term "bridge" is used here in a slightly generalized way meaning any Jordan arc β outside K joining two points on K.

THE LAPLACE EQUATION IN SPACE

BY EDWARD KASNER AND JOHN DE CICCO

DEPARTMENTS OF MATHEMATICS, COLUMBIA UNIVERSITY AND ILLINOIS INSTITUTE OF TECHNOLOGY

Communicated June 29, 1945

1. A function $h(x, y)$ is harmonic if it satisfies the Laplace equation in two dimensions $h_{xx} + h_{yy} = 0$. The one-parameter family of curves $h(x, y) = c$ is called an isothermal family and c is said to be the isothermal parameter. A family of curves $f(x, y) = c$ is isothermal if and only if f is a function of a harmonic function, and, therefore, if f satisfies a certain partial differential equation of third order.

Sophus Lie obtained the first intrinsic characterization of the set of isothermal families, which may be stated in the following form. A family of ∞^1 curves is isothermal if and only if the inclination θ is a harmonic function of (x, y). From this it follows that the angle between any two isothermal families is a harmonic function of (x, y).

The authors have investigated all possible types of transformations preserving the class of isothermal families. We have found extensive classes of differential element transformations converting isothermal families into isothermal families. For the case of lineal-element transformations, the following result has been proved. The group of lineal-element transformations preserving the class of isothermal families is: $X = \phi(x, y)$, $Y = \psi(x, y)$, $\Theta = a\theta + h(x, y)$, where ϕ and ψ satisfy the direct or reverse Cauchy-Riemann equations: $\phi_x = \pm\psi_y$, $\phi_y = \mp\psi_x$, a is a non-zero constant and h is any harmonic function of (x, y). By this, we establish the theorem that the only point transformations which send every isothermal family into an isothermal family, are the conformal. (This is more general and more difficult than the obvious problem of finding the auto-transformations of the Laplace equation.)

2. In our present work, we shall study the case in three dimensions, where the situation is entirely different both analytically and geometrically. A function $h(x, y, z)$ is harmonic if it satisfies the Laplace equation $h_{xx} + h_{yy} + h_{zz} = 0$. The ∞^1 surfaces $h(x, y, z) = c$ is called an isothermal family and c is said to be the isothermal parameter. A family of surfaces $f(x, y, z) = c$ is isothermal if and only if f is a function of a harmonic function, and, therefore, if and only if f satisfies two partial differential equations of third order. Now we state our new theorems.

THEOREM 1. *If h is any harmonic function of three variables (x, y, z) with $h_x \neq 0$, $h_y \neq 0$, $h_z \neq 0$, then a function $f = f(h_x, h_y, h_z)$ is harmonic for any arbitrary h only when f is linear integral in the three arguments.*

This theorem can be extended to partial derivatives of h of any order,

that is, the function f must be linear integral in all the partial derivatives. There is a similar theorem which is valid for any dimension $n \geqq 3$.

However, for $n = 2$ the situation is quite different. If h is a harmonic function of two variables (x, y), with $h_x \neq 0$, $h_y \neq 0$, then $f = f(h_x, h_y)$ is harmonic for any h if and only if f is harmonic in the two arguments. Thus arc tan h_y/h_x, $\log (h_x^2 + h_y^2)$, $e^{h_x} \cos h_y$, $h_x^2 - h_y^2$ are all harmonic functions for any harmonic function h, although non-linear.

3. THEOREM 2. *Lie's characterization of isothermal families in the plane is not valid in space. That is, the angle between any isothermal system of surfaces and a set of parallel planes is not necessarily a harmonic function.*

This result can be deduced from Theorem 1. For if $h(x, y, z)$ is a harmonic function, then the angle γ between the isothermal system of surfaces $h(x, y, z) =$ const. and the pencil of parallel planes $z =$ const., is a complicated expression involving the partial derivatives of h. By Theorem 1, this angle γ cannot be a harmonic function for every h.

4. THEOREM 3. *The system of simultaneous partial differential equations of first order:* $\partial z/\partial x = p(x, y, z)$, $\partial z/\partial y = q(x, y, z)$, *possesses as solution an isothermal family of surfaces if and only if p and q satisfy the set of three partial differential equations of second order.*

$$\begin{aligned} p_y + qp_z &= q_x + pq_z, \\ {}^1/_2(1 + p^2 + q^2)(p_{xx} + p_{yy} + p_{zz}) &= (p_yq_z - p_zq_y) + p(p_x^2 + p_y^2 + p_z^2 + \\ &\quad p_xq_y - p_yq_x) + q(p_xq_x + p_yq_y + p_zq_z), \\ {}^1/_2(1 + p^2 + q^2)(q_{xx} + q_{yy} + q_{zz}) &= (p_zq_x - p_xq_z) + q(q_x^2 + q_y^2 + q_z^2 + \\ &\quad p_xq_y - p_yq_x) + p(p_xq_x + p_yq_y + p_zq_z). \quad \text{(I)} \end{aligned}$$

The first of these conditions is the condition of integrability.

As an application of Theorem 3, the following result may be established. The only families of ∞^1 planes which are isothermal are the pencils of planes. For straight lines the analogous result was proved by Lagrange. The proof in space is difficult.

5. THEOREM 4. *The only point transformations converting every isothermal system of surfaces into an isothermal system are those of the Liouville inversive group.*

This proposition is established with the aid of Liouville's theorem which states that the conformal group of space is exactly the inversive group of ten parameters.

6. At this stage, let us consider transformations of surface elements (x, y, z, p, q). The Lie contact group satisfies the conditions

$$\left.\begin{aligned} Z_x + pZ_z &= P(X_x + pX_z) + Q(Y_x + pY_z), Z_p = PX_p + QY_p, \\ Z_y + qZ_z &= P(X_y + qX_z) + Q(Y_y + qY_z), Z_q = PX_q + QY_q. \end{aligned}\right\} \text{(L)}$$

The Lie group has the property that it converts every union, that is, any

double series: $z = z(x, y)$, $p = p(x, y)$, $q = q(x, y)$, with the property that $dz = pdx + qdy$, into a union.

Kasner studied transformations of surface elements which carry every integrable field of planar elements, that is, $p = p(x, y, z)$, $q = q(x, y, z)$, which satisfy the condition $p_y + qp_z = q_x + pq_z$, into an integrable field. It is not assumed that a union of an integrable field is carried into the corresponding union of the associated integrable field. Kasner proved that *this group of transformations is identical with the Lie contact group.*

7. By means of Kasner's theorem, it follows that any transformation of surface elements which carries every isothermal field into an isothermal field must be a contact transformation. Then the following result may be established.

THEOREM 5. *The only transformations of surface elements which send every isothermal field into an isothermal field are those of the Liouville inversive group.*

Thus in space the only possible transformations preserving the isothermal character must be point transformations generated by inversions with respect to a sphere.

[1] Kasner, *Differential Geometric Aspects of Dynamics*, Princeton Colloquium Lectures, 1913, 1934.

[2] Kasner, "A Characteristic Property of Isothermal Families," *Mathematische Annalen*, 59, 352–354 (1904).

[3] De Cicco, "New Proofs of the Theorems of Beltrami and Kasner on Linear Families of Curves," *Bull. Am. Math. Soc.*, 49, 407–412 (1943).

[4] De Cicco, "The Two Conformal Covariants of a Field," *Revista de Matemáticas de Tucumán*, 2, 59–66 (1941).

[5] Kasner, "Lineal-element Transformations Which Preserve the Isothermal Character," these PROCEEDINGS, 27, 406–412 (1941).

[6] Kasner, "Transformation Theory of Isothermal Families and Certain Related Trajectories," *Revista de Matemáticas y Física Teórica de la Universidad de Tucumán*, 2, 17–24 (1941), Argentina.

[7] Kasner and De Cicco, "Generalized Transformation Theory of Isothermal and Dual Families," these PROCEEDINGS, 28, 52–55 (1942).

[8] Kasner and De Cicco, "An Extensive Class of Transformations of Isothermal Families," *Revista de Matemáticas y Física Teórica de la Universidad de Tucumán*, 3, 271–282 (1942).

[9] Kasner and De Cicco, "Transformation Theory of Isogonal Trajectories of Isothermal Families," these PROCEEDINGS, 28, 328–333 (1942).

[10] Kasner, "Geometric Properties of Isothermal Families," *Facultad de Ciencias Matematicas de la Universidad Nacional del Litoral*, 5, 1–10 (1943), Rosario, Argentina.

[11] Kasner, "Lineal-Element Transformations of Space for Which Normal Congruences of Curves are Converted into Normal Congruences," *Duke Mathematical Journal*, 5, 72–83 (1939).

ALGEBRAIC CURVES, SYMMETRIES AND SATELLITES

By Edward Kasner

Department of Mathematics, Columbia University

Communicated June 29, 1945

Schwarz defined symmetry with respect to any real analytic base curve. This is known as Schwarzian reflection and is used extensively in the theory of analytic prolongation of functions of a complex variable. The author defined the process geometrically and intrinsically in the Proceedings of the Cambridge International Congress (1912) and called it *conformal symmetry*. If the base curve is denoted by C, then the image of any point P of the plane is obtained in the following manner. Through P, draw the two minimal lines, which will intersect C, respectively, in the two points Q_1 and Q_2. The remaining minimal lines passing through Q_1 and Q_2 will intersect in a point P'. This point P' is the image of the point P with respect to the base curve C. *Conformal symmetry is the only reverse conformality which leaves the individual points of C fixed. No other reverse conformalities of period two can exist.*

In the present paper, I wish to discuss the case where the base curve C is a general algebraic curve. We thus operate in the complete complex (four-dimensional) plane. Schwarz was considering only the real (two-dimensional) gaussian plane.

In the Schwarzian theory, the transformation is one valued because only the local neighborhood of the base curve C is considered. But in the present paper, since we are considering the total plane, the transformation T is many valued. If the algebraic curve C is of degree n, then the transformation T is always algebraic and is, in general, of degree n^2.

If, in particular, the base curve C is a conic, the transformation T will convert, in general, one point P into four points P'. But if the conic is a circle, we obtain ordinary inversion, which is of course one valued. There are mixed imaginary cases of conics where the degree is neither four nor one but is actually two.

Usually we say that the image of the base curve C with respect to itself is C. But in the algebraic case we are now considering, this is no longer exactly true. The complete image of C consists partly of C and partly of a new curve which we define as the *satellite* of C and denote by S. The image of a random curve C' of degree n with respect to C is, in general, of degree n^3. But the image of C with respect to itself is reducible, and we find that one branch is the curve C counted a certain number of times, and the new satellite curve S.

In the case of a conic C, the satellite S is another conic. The new conic S is confocal with the old conic C, and has uniquely determined diameters.

If C is a rectangular hyperbola, the satellite conic S is identical with C. Of course, the circle has no satellite.

If we consider the totality of ∞^5 conics of the plane, the induced transformation from each conic to its satellite is of the fifth degree (in the coefficients of the conic). This induced transformation is neither a point transformation nor a contact transformation. Of course, all the contact transformations carrying the set of ∞^5 conics into itself form merely the total projective group consisting of collineations and correlations.

THEOREM 1. *The degree of the satellite S of a general algebraic curve C of degree n is $n(n - 1)^2$.*

This is true only in general since actually the degree of the satellite may in special cases be lower. From this formula, we see that the satellite of a conic is a conic; but the satellite S of a cubic curve C is, in general, an algebraic curve of degree 12.

A noteworthy case where the degree of the satellite is lower is stated in the following result.

THEOREM 2. *The satellite S of every algebraic potential curve C is the curve C itself.*

The curve $\phi(x, y) = 0$, where ϕ is a polynomial of degree n in (x, y), is called a potential curve if ϕ is harmonic, that is, $\phi_{xx} + \phi_{yy} = 0$. In this case, although the transformation T associated with the curve is of degree n^2, reducibility studies show that the satellite is of exceptionally low degree. When the degree is 2, we obtain the previously noted example of the equilateral hyperbola.

We have already observed that if the degree of C is n, the degree of T is n^2. What then is the degree of T^2? We might expect it to be of degree n^4. But we prove the following proposition (on account of reducibility).

THEOREM 3. *The degree of T^2 is $n^2(n - 1)^2$.*

All this theory depends on the fact that the transformations are algebraic multivalued, and that reducibility phenomena occur.

THEOREM 4. *If the base curve C is a conic, T is of degree 4 as already noted; and we now state that T^2 is also of degree 4, and all the iterations of T give transformations of degree 4.*

It thus turns out that the various powers of T form a discontinuous set. This set has the combinatorial property but no power of T is the identity. Thus this set, although closed, cannot be a group.

An interesting case occurs in the imaginary domain when the base conic C goes through only one circular point at infinity. The transformation T is then of degree 2. We find that all the powers of T are of degree 2. It results that $T^3 = T$ but no power of T is the identity. Thus, in this case, the complete set of all powers of T consists essentially of only the two transformations T and T^2.

We also have obtained results on *related families* of curves $\phi(x, y) = C$

their satellites, and set of symmetries, and a remarkable *induced transformation* of lower degree than the symmetries. For the case of conics the induced transformation is linear, and is in fact an affinity. This affinity enables us to construct the satellite conic easily in the real domain.

[1] Kasner, "Algebraic Potential Curves," *Bull. Am. Math. Soc.*, 1901.

[2] Kasner, "Conformal Geometry," *Proc. Fifth Internat. Cong. of Math.*, 2, 81 (1912).

[3] Kasner, "Geometry of Conformal Symmetry (Schwarzian Reflection)," *Ann. Math.*, 38, 873 (1937).

[4] Comenetz, "Conformal Geometry on a Surface," *Ibid.*, 39, 863 (1938).

I wish to thank J. De Cicco, Aida Kailish, F. Supnik and C. Vlucay for assistance in elaborating the theory of satellites.

DIFFERENTIAL EQUATIONS IN FRÉCHET DIFFERENTIALS OCCURRING IN INTEGRAL EQUATIONS

BY ARISTOTLE D. MICHAL

CALIFORNIA INSTITUTE OF TECHNOLOGY

Communicated July 5, 1945

Introduction.—This paper is concerned with the study of the resolvent kernels and solutions of Volterra and Fredholm integral equations as functionals of the given kernels. A study is made of the completely integrable equations in Fréchet differentials characterizing these functionals of the kernels. The investigations have immediate application to the problem of obtaining approximations to the resolvent kernels and the solutions of integral equations with precise estimates of the errors. The above results are applied to differential corrections for some differential equations. *As an important by-product we obtain a solution of a century-old problem in non-commutative analysis.* We assume that the reader is conversant with the fundamental theorems on Fréchet differentials of functionals.

1. *Resolvent Kernels and Solutions of Volterra Integral Equations as Functionals of the Kernels.*—Consider the Volterra integral equation of the second kind

$$f(x) = y(x) + \int_a^x K(x, \xi)y(\xi)d\xi (T: a \leq \xi \leq x \leq b) \qquad (1)$$

with continuous kernel in T and continuous $f(x)$ in (a, b). The well-known unique continuous solution $y(x)$ of equation (1) in the closed interval (a, b) is

$$y(x) = f(x) + \int_a^x k(x, \xi)f(\xi)d\xi, \qquad (2)$$

where the resolvent kernel $k(x, \xi)$ has the expansion

$$k(x, \xi) = -K(x, \xi) + K^2 - K^3 + \ldots + \ldots \qquad (3)$$

in terms of the composition powers of the kernel $K(x, \xi)$:

$$K^2 = \int_\xi^x K(x, s)K(s, \xi)ds, \qquad K^{i+1} = \int_\xi^x K(x, s)K^i(s, \xi)ds. \quad (4)$$

We shall write $k(x, \xi)$ as $k[K]$ to show its functional dependence on the kernel $K(x, \xi)$. Since $k[K]$ is an entire analytic functional of $K(x, \xi)$, it follows from a theorem[1] of the author on the Fréchet differentiability of regular power series in normed linear spaces that the Fréchet differential $\delta k[K]$ of $k[K]$ with increment δK exists for each continuous $K(x, \xi)$ in T and is given by a term by term Fréchet differentiation of (3). The result written in terms of composition products is

$$\delta k[K] = -(\delta K + k\delta K + \delta K k + k\delta K k). \quad (5)$$

For example $k\delta K$ stands for $\int_\xi^x k(x, s)\delta K(s, \xi)ds$. In calculating this Fréchet differential and in considering $k[K]$ as an analytical functional of $K(x, \xi)$, we understand that the independent and dependent variables in $k[K]$ are in the complete normed linear space (a Banach space) of all continuous $K(x, \xi)$ in T with norm defined by

$$||K|| = \max_{x,\, \xi \epsilon T} |K(x, \xi)|. \quad (6)$$

In (5), K as well as the increment δK are arbitrary continuous functions of x and ξ in T.

Another way of obtaining formula (5) is to use the existence of the Fréchet differential $\delta k[K]$, guaranteed by the author's theorem on regular power series in normed linear spaces, and then to *compute* it from either of the classical resolvent relations of integral equation theory:

$$K + k + Kk = 0, \qquad k + K + kK = 0. \quad (7)$$

In fact, if we take the Fréchet differential of both sides of the first of these identities, we obtain

$$\delta K + \delta k + \delta K k + K\delta k = 0. \quad (8)$$

It can be shown that the unique continuous solution of this integral equation (8) with δk as unknown is given precisely by (5).

Since δk exists given by (5), the Fréchet differential of the solution $y[K/x]$ of the integral equation (1) considered as a functional of the kernel $K(x, \xi)$ can be computed from (2). A simpler derivation is the following which does not use the result (5) but merely the existence of δk and hence of $\delta y[K/x]$ from (2). Take the Fréchet differential of both sides of (1) and obtain

$$0 = \delta y[K/x] + \int_a^x \delta K(x, \xi)y[K/\xi]d\xi + \int_a^x K(x, \xi)\delta y[K/\xi]d\xi.$$

If we solve this Volterra integral equation of the second kind for $\delta y[K/\xi]$, we obtain a result which can be reduced to the form

$$\delta y[K/x] = -\int_a^x \left\{\delta K(x, \xi) + \int_\xi^x k(x, s)\delta K(s, \xi)ds\right\} y[K/\xi]d\xi. \quad (9)$$

In the above discussion it is understood that the Fréchet differential $\delta y[K/x]$ is defined for a functional $y[K/x]$ with values in the well-known Banach space of continuous functions of x. Results (5) and (9) for one Volterra integral equation can, by an evident change of interpretation of notation, be shown to hold for a system of n Volterra integral equations of the second kind in n unknowns. If the system of equations is written as (1) with $f(x)$ and $y(x)$ as column matrices of functions of x and $K(x, s)$ as the n-rowed square matrix of the n^2 kernels, then in (2), $k(x, \xi)$ will stand for the n-rowed square matrix of the n^2 resolvent kernels. In (5) products are to be interpreted as combined matric and integral composition products of the first kind; for example, $k\delta K$ will stand for $\int_\xi^x k^i_\alpha(x, s)\delta K^\alpha_j(s, \xi)ds$ with the summation convention operating on indices. In (9), products of functions will be matric products.

2. The Completely Integrable Total Differential Equations in Fréchet Differentials of Functionals.—Let us now return to the one Volterra integral equation (1) and inquire into the properties of the functional equations (5) and (9) satisfied, respectively, by the resolvent kernel and the solution of (1) as functionals of the kernel $K(x, \xi)$.

It can be shown from (5) with the aid of the known theorems on differentials and a mathematical induction that all successive Fréchet differentials of the resolvent kernel $k[K]$ exist at any chosen continuous kernel $K(x, \xi)$ in T. If $\delta_1 K(x, \xi)$, $\delta_2 K(x, \xi)$, $\ldots$, $\delta_n K(x, \xi)$ are the increment functions for the nth successive Fréchet differential $\delta_n\delta_{n-1}\ldots\delta_1 k[K]$, we find that

$$\left.\begin{array}{l}\delta_n\delta_{n-1}\ldots\delta_1 k[K] = (-1)^n P_{1,2,\ldots,n}\{[\delta_1 K + k\delta_1 K)(\delta_2 K + \\ \quad k\delta_2 K)\ldots(\delta_n K + k\delta_n K) + (\delta_1 K + k\delta_1 K)(\delta_2 K + k\delta_2 K)\ldots \\ \qquad\qquad (\delta_n K + k\delta_n K)k\}(n = 1, 2, 3, \ldots),\end{array}\right\} \tag{10}$$

where $P_{1,2}\ldots,n$ stands for the sum of $n!$ terms obtained from (and including) the bracket by a permutation of the integers $1, 2, \ldots, n$. If all increments are equal, the nth Fréchet differential $\delta^n k[K]$ is given by

$$\delta^n k[K] = (-1)^n n!\{(\delta K + k\delta K)^n + (\delta K + k\delta K)^n k\}(n = 1, 2, 3, \ldots). \tag{11}$$

The total differential equations (5) and (9) are *completely integrable* since the condition for integrability is identically satisfied in all continuous K, k, $\delta_1 K$, $\delta_2 K$ (considered as independent variables) in T. Although equations (5) and (9) are completely integrable for all continuous kernels K in T, the existence and uniqueness theorems[2] in normed linear spaces of Michal and Elconin are not strong enough to insure the unicity of the solution taking on arbitrary initial conditions. We can, however, give an argument which will furnish uniqueness theorems for functional equations (5) and (9). In fact let $\lambda[K]$ satisfy the following equation in Fréchet differentials $\delta\lambda[K]$ for all kernels $K(x, \xi)$ continuous in T:

$$\delta\lambda[K] = -(\delta K + \lambda\delta K + \delta K\lambda + \lambda\delta K\lambda), \qquad (12)$$

and let it take on the initial condition $\lambda[0] = 0$, the identically vanishing function in T. We shall find the most general such entire analytical functional.[3] Assume then that

$$\lambda[K] = \sum_{i=1}^{\infty} \lambda_i[K] \qquad (13)$$

so that $\lambda_i[K]$ is a homogeneous functional polynomial of degree i. It follows from R. S. Martin's contributions to abstract polynomials[4] that to each homogeneous polynomial $\lambda_i[K]$ there exists a unique polar, i.e., a completely symmetric multilinear functional $\Lambda_i[K_1, K_2, \ldots, K_i]$ such that $\Lambda_i[K, K, \ldots, K] = \lambda_i[K]$. It is readily shown that the Fréchet differential $\delta\lambda_i[K]$ is given by

$$\delta\lambda_i[K] = i\Lambda_i[K, K, \ldots, K, \delta K]. \qquad (14)$$

By the author's theorem[1] on the Fréchet differentiability of power series in normed linear spaces, we have

$$\delta\lambda[K] = \sum_{i=1}^{\infty} i\Lambda_i[K, K, \ldots, \delta K] \qquad (15)$$

for all $K(x, \xi)$ continuous in T.

On using (13), (14) and (15) in (12), one can show that $\lambda_1[K] = -K$, $\lambda_2[K] = K^2$ and that the following recurrence relations hold for $n \geq 2$:

$$(n + 1)\Lambda_{n+1}[K, K, \ldots, K, \delta K] = -\lambda_n[K]\delta K - \delta K\lambda_n[K] - \sum_{\substack{i+j=n \\ i,j\geq 1}} \lambda_i[K]\delta K\lambda_j[K].$$

From these results one finds that

$$\lambda[K] = -K + K^2 - K^3 + \ldots,$$

and hence from our earlier results on resolvent kernels the fundamental THEOREM 1. *The completely integrable non-linear total differentia system in Fréchet differentials*

$$\left.\begin{aligned} \delta\lambda[K] &= -(\delta K + \lambda\delta K + \delta K\lambda + \lambda\delta K\lambda) \\ \lambda[0] &= 0 \end{aligned}\right\} \qquad (16)$$

has a unique entire analytic functional solution given by

$$\lambda[K] = -K + K^2 - K^3 + \ldots,$$

the resolvent kernel of the arbitrary continuous kernel $K(x, \xi)$ in T.

By similar methods one can prove the following theorem.

THEOREM 2. *The completely integrable linear total differential system in Fréchet differentials, with $k[K/x, \xi]$ the resolvent kernel of $K(x, \xi)$,*

$$\left.\begin{aligned} \delta Z[K/x] &= -\int_a^x \left\{ \delta K(x, \xi) + \int_\xi^x k[K/x, s]\delta K(s, \xi)ds \right\} Z[K/\xi]d\xi \\ Z[0/x] &= f(x) \end{aligned}\right\} \quad (17)$$

has a unique entire analytic functional solution given by

$$Z[K/x] = f(x) + \int_a^x k[K/x, \xi]f(\xi)d\xi, \quad (18)$$

the solution of the Volterra integral equation (1).

In this theorem we understand that the definitions of Fréchet differentials and analytic functionals for $Z[K/x]$ are given with the independent variable $K(x, \xi)$ ranging over the Banach space of the previous theorem while the value space is the Banach space of functions of x continuous in the interval $a \leq x \leq b$ and having as norm the maximum of the absolute value of the function over the interval. The details of proof are facilitated by writing the differential system as

$$\begin{aligned} \delta Z[K] &= -(\delta K + k[K]\delta K)\cdot Z[K] \\ Z[0] &= f \end{aligned}$$

and by writing $A\cdot w$ for the bilinear functional $\int_a^x A(x, \xi)w(\xi)d\xi$.

A more general theorem can be proved by different methods on making use of Theorem 1 and the result $\delta(Z[K] + K\cdot Z[K]) = 0$.

THEOREM 3. *There is one and only one solution of the completely integrable system (17) in Fréchet differentials. It is given by (18).*

3. *Simple Illustrative Applications to Ordinary Differential Equations.*—The differential system ($P(x)$ and $Q(x)$ continuous, say, in (a, b))

$$\frac{dy(x)}{dx} + P(x)y(x) = Q(x), \qquad y(a) = y_0 \quad (19)$$

is equivalent to an integral equation (1) with kernel $K(x, \xi) = P(\xi)$ and known function $f(x) = y_0 + \int_a^x Q(s)ds$. With the aid of Theorem 2 one can show that the solution $y[P/x]$ of the system (19) as a functional of $P(x)$ is the unique entire analytic functional solution of the total differential system in Fréchet differentials.

$$\left.\begin{aligned} \delta y[P/x] &= -\int_a^x e^{-\int_s^x P(u)du} y[P/s]\delta P(s)ds \\ y[0/x] &= y_0 + \int_a^x Q(t)dt. \end{aligned}\right\} \quad (20)$$

As another simple example we take the differential system

$$\frac{d^2y(x)}{dx^2} + P(x)y(x) = 0, \qquad y(a) = y_0, \qquad \left.\frac{dy(x)}{dx}\right]_{x=a} = y_1$$
$$(P(x) \text{ continuous in } (a, b)). \quad (21)$$

This is equivalent to an integral equation (1) with kernel $K(x, \xi) = (x - \xi) P(\xi)$ and known function $f(x) = y_0 + y_1(x - a)$. Again with the aid of Theorem 2 one can show that the solution $y[P/x]$ of the system (21) as a functional of $P(x)$ is the unique entire analytic functional solution of the differential system in Fréchet differentials

$$\left.\begin{aligned} \delta y[P/x] &= \int_a^x \mu[P/x, \xi] y[P/\xi] \delta P(\xi) d\xi \\ y[0/x] &= y_0 + y_1(x - a), \end{aligned}\right\} \tag{22}$$

where

$$\mu[P/x, \xi] = -(x - \xi) - \int_\xi^x F[P/x, \eta] (\eta - \xi) d\eta \tag{23}$$

and $F[P/x, \eta]$ is the resolvent kernel of the kernel $K(x, \eta) = (x - \eta)P(\eta)$.

4. *The Matric Exponential in Non-Commutative Analysis.*—The results of Theorem 1, Theorem 2 and Theorem 3 can easily be extended to the case of a system of n Volterra integral equations. We shall apply such a generalized Theorem 2 to the system of differential equations with constant coefficients

$$\frac{dZ^i(x)}{dx} = a^i_j Z^j \qquad Z^i(0) = Z^i_0. \tag{24}$$

This is equivalent to the matric Volterra integral equation

$$Z_0 = Z(x) + \int_0^x K(x, \xi) Z(\xi) d\xi \tag{25}$$

where $K(x, \xi) = -A = -||a^i_j||$, the matrix of the constants a^i_j.

If we write the solution of (25) as $Z[K/x]$, we see that

$$Z[-A/1] = e^A Z_0, \tag{26}$$

where e^A is the *matric exponential.* With the aid of the generalized Theorem 2, we can show that

$$\begin{aligned} \delta(e^A Z_0) &= \int_0^1 \{\delta A + \int_\xi^1 A e^{(1-s)A} \delta A ds\} e^{\xi A} Z_0 d\xi, \\ k[0] &= 0. \end{aligned}$$

This leads to the following important result giving the functional equation satisfied by the matric exponential.

THEOREM 4. *The matric exponential function $Z(A) = e^A$, satisfies the folowing system in Fréchet differentials:*[5]

$$\left.\begin{aligned} \delta Z(A) &= \int_0^1 Z((1 - \xi)A) \delta A Z(\xi A) d\xi \\ Z(0) &= I, \textit{ the unit matrix.} \end{aligned}\right\} \tag{27}$$

5. *Resolvent Kernels and Solutions of Fredholm Integral Equations as Functionals of the Kernels.*—The set Δ of kernels $K(x, \xi)$ continuous in the square $S: a \leq x, \xi \leq b$ and with non-vanishing Fredholm determinants forms an open set in the Banach space of all continuous $K(x, \xi)$ in S. If we now write a composition of the second kind (constant limits) merely as a product, the following result can be proved.

THEOREM 5. *The Fréchet differential of the resolvent kernel* $k[K]$ *of a kernel* $K(x, \xi)$ *exists at each* $K(x, \xi) \epsilon \Delta$, *and for all* $K(x, \xi) \epsilon \Delta$ *it satisfies the differential system in Fréchet differentials*

$$\begin{aligned} \delta k[K] &= -(\delta K + k\delta K + \delta K k + k\delta K k), \\ k\,[0] &= 0. \end{aligned}$$

The proof of this theorem[6] differs from the power series proof for the corresponding result in Volterra integral equations.

The correspondent of Theorem 3 can be proved by entirely analogous methods. We shall state it as

THEOREM 6. *There is one and only one solution of the completely integrable system in Fréchet differentials*

$$\begin{aligned} \delta Z[K/x] &= -\int_a^b \{\delta K(x, \xi) + \int_a^b k[K/x, s]\delta K(s, \xi)ds\} Z[K/\xi]d\xi, \\ Z[0/x] &= f(x), \qquad (K \epsilon \Delta) \end{aligned}$$

where $k[K/x, s]$ *is the resolvent kernel of* $K(x, \xi)$. *It is given by the solution of the Fredholm integral equation with kernel* $K(x, \xi)$ *and known function* $f(x)$.

[1] Michal, A. D., *The Fréchet Differentials of Regular Power Series in Normed Linear Spaces*. To be published elsewhere.

[2] Michal, A. D., and Elconin, V., "Completely Integrable Differential Equations in Abstract Spaces," *Acta Mathematica*, **68**, 71–107 (1937).

[3] In Banach spaces, $\sum_{i=1}^{\infty} \lambda_i[K]$ defines an entire analytic function if the corresponding real power series $\sum_{i=1}^{\infty} m_i x^i$ converges for all real x. By m_i we mean the modulus of the homogeneous polynomial $\lambda_i[K]$. For analytic functions in Banach spaces see Michal, A. D., and Martin, R. S., "Some Expansions in Vector Space," *Jour. Math. Pures et Appl.*, **13**, 69–91 (1934). Presented to the American Mathematical Society, Sept., 1932. See also Martin, R. S., *Contributions to the Theory of Functionals*, a California Institute of Technology Ph.D. thesis (unpublished), 1932, written under the direction of the present author. Some portions of this thesis have trickled into the later mathematical literature.

[4] Martin, R. S., *Contributions to the Theory of Functionals*, loc. cit. For essentially Martin's proof of the existence and uniqueness of a polar see also Taylor, A. E., "Additions to the Theory of Polynomials in Normed Linear Spaces," *Tôhoku Math. Jour.*, **44**, 302–318 (1938), where other references are given. See also Van Der Lijn, *Bulletin des Sciences Math.* (1940).

[5] The system (27) in Fréchet differentials is not of the type for which existence theorems were given by Michal and Elconin, loc. cit. For some further results on the matric exponential, the reader is referred to Michal, A. D., *The Total Differential Equation for the Exponential Function in Non-Commutative Normed Linear Rings* (to be published). For some applications of the matric exponential to vibration problems, the reader is referred to Michal, A. D., *Matrix and Tensor Calculus with Applications to Mechanics, Elasticity, and Aeronautics* (GALCIT series of John Wiley and Sons, in press).

[6] The functional of ξ defined by $L(K, \xi) = \xi + \xi K$ is a solvable linear functional of ξ with inverse $M(K, \xi) = \xi + \xi k[K]$. Hence by a known result (Michal and Elconin, loc. cit.) the Fréchet differential of $\xi k[K]$ exists for each $K(x, \xi)\epsilon\Delta$. With this result one can complete the proof of Theorem 5 without much difficulty. This same type of proof can also be used in proving the corresponding result (5) for Volterra integral equations.

ON THE PROJECTIVE THEORY OF TWO DIMENSIONAL RIEMANN SPACES

BY T. Y. THOMAS

DEPARTMENT OF MATHEMATICS, INDIANA UNIVERSITY

Communicated June 25, 1945

The only significant projective invariant which has been defined in the strict domain of the projective space of paths is the projective curvature tensor discovered by Weyl[1] in 1921. In the following note we show the existence of a simple *vector* invariant of the two-dimensional projective Riemann space (defined by the totality of Riemann metrics which yield the same system of geodesics). The vanishing of this vector constitutes the necessary and sufficient condition for the space to be of constant curvature in the projective sense.

The construction of the above vector invariant follows readily from equations which have already appeared in the literature.[2] Thus if $g_{\alpha\beta}(x)$ and $\bar{g}_{\alpha\beta}(x)$ are the components of metric tensors yielding the same system of geodesics we must have

$$\bar{\Gamma}^{\alpha}_{\beta\gamma} = \Gamma^{\alpha}_{\beta\gamma} + \delta^{\alpha}_{\beta}\phi_{\gamma} + \delta^{\alpha}_{\gamma}\phi_{\beta}, \tag{1}$$

where the Γ's and $\bar{\Gamma}$'s are the Christoffel symbols determined by these tensors and the ϕ's are the components of a covariant vector. Equations (1) are completely equivalent to the system

$$\bar{g}_{\alpha\beta,\gamma} = \bar{g}_{\alpha\gamma}\phi_{\beta} + \bar{g}_{\beta\gamma}\phi_{\alpha} + 2\bar{g}_{\alpha\beta}\phi_{\gamma}, \tag{2}$$

in which the comma denotes covariant differentiation based on the metric defined by the $g_{\alpha\beta}$. Contraction of indices in (1) shows that the vector having the components ϕ_α is the gradient of a scalar function ϕ given by

$$\phi = \frac{1}{2(n+1)} \log\left(\frac{\bar{g}}{g}\right),$$

where g and $\bar{g}$ denote the determinants formed from the components $g_{\alpha\beta}$ and $\bar{g}_{\alpha\beta}$, respectively, and n is the dimensionality of the space. Substitution of the components $\bar{\Gamma}^{\alpha}_{\beta\gamma}$ from (1) into the expression for the components of the ordinary curvature tensor, followed by contraction of indices, leads to the system

$$\bar{B}_{\alpha\beta} = B_{\alpha\beta} + (n-1)[\phi_\alpha\phi_\beta - \phi_{\alpha\beta}], \tag{3}$$

where the $B_{\alpha\beta}$ and $\bar{B}_{\alpha\beta}$ are the components of the contracted curvature tensors and the $\phi_{\alpha\beta}$ are the components of the second covariant derivative of the scalar ϕ.

We now suppose $n = 2$. Then the curvature tensor components $B^{\alpha}_{\beta\gamma\delta}$ can be written

$$B^{\alpha}_{\beta\gamma\delta} = g^{\alpha\sigma}B_{\sigma\beta\gamma\delta} = g^{\alpha\sigma}(g_{\sigma\delta}g_{\beta\gamma} - g_{\sigma\gamma}g_{\beta\delta})K = (\delta^{\alpha}_{\delta}g_{\beta\gamma} - \delta^{\alpha}_{\gamma}g_{\beta\delta})K, \quad (4)$$

where K is the Gaussian curvature. Putting $\alpha = \delta$ in these equations and summing we find $B_{\beta\gamma} = Kg_{\beta\gamma}$. Hence (3) becomes

$$\phi_{\alpha\beta} = \phi_{\alpha}\phi_{\beta} + Kg_{\alpha\beta} - \bar{K}\bar{g}_{\alpha\beta}. \quad (5)$$

Now differentiate (5) covariantly with respect to x^{γ}, interchange the indices β and γ and subtract. This gives

$$\phi_{\alpha\beta,\gamma} - \phi_{\alpha\gamma,\beta} = \phi_{\alpha\gamma}\phi_{\beta} - \phi_{\alpha\beta}\phi_{\gamma} + K_{\gamma}g_{\alpha\beta} - K_{\beta}g_{\alpha\gamma} - \bar{K}_{\gamma}\bar{g}_{\alpha\beta} + \bar{K}_{\beta}\bar{g}_{\alpha\gamma} - \bar{K}\bar{g}_{\alpha\beta,\gamma} + \bar{K}\bar{g}_{\alpha\gamma,\beta}. \quad (6)$$

For the left member of these equations we have

$$\phi_{\alpha\beta,\gamma} - \phi_{\alpha\gamma,\beta} = -\phi_{\sigma}B^{\sigma}_{\alpha\beta\gamma} = (\phi_{\beta}g_{\alpha\gamma} - \phi_{\gamma}g_{\alpha\beta})K.$$

Making this substitution, and the substitutions (2) and (5) in the right member of (6), the resulting expression is seen to reduce to

$$\bar{K}_{\beta}\bar{g}_{\alpha\gamma} - \bar{K}_{\gamma}\bar{g}_{\alpha\beta} = K_{\beta}g_{\alpha\gamma} - K_{\gamma}g_{\alpha\beta}. \quad (7)$$

It follows from (7) that the quantities $K_{\beta}g_{\alpha\gamma} - K_{\gamma}g_{\alpha\beta}$ are the components of a projective Riemann tensor. Since these quantities are skew symmetric in the indices β and γ this tensor involves at most two independent components. This suggests that the tensor is equivalent to a vector invariant. To show that such is the case we proceed as follows: make the substitutions

$$g_{11} = gg^{22},\ g_{12} = -gg^{12},\ g_{22} = gg^{11}$$

and similar substitutions in the quantities $\bar{g}_{\alpha\beta}$ in the equations (7). These equations then yield

$$\bar{g}\bar{g}^{\alpha\beta}\bar{K}_{\beta} = gg^{\alpha\beta}K_{\beta}. \quad (8)$$

Hence the quantities $gg^{\alpha\beta}K_{\beta}$ are the components of a relative projective vector of weight two of the two dimensional Riemann space. The vanishing of this vector implies $K_{\alpha} = 0$ and conversely. Thus the only spaces whose geodesics are identical with the geodesics of a space of constant curvature are spaces of constant curvature (Beltrami). Moreover if $K =$ const., the constant $\bar{K}$ can be chosen arbitrarily; this follows readily since under these conditions the system consisting of (2) and (5) is completely integrable. The totality of Riemann spaces of constant curvature having the same geodesics can conveniently be described as a projective Riemann space of constant curvature. Hence the condition for a two-dimensional space to be of constant curvature in the projective sense (projective Rie-

mann space of constant curvature) is the vanishing of the above projective vector invariant.

It follows from (8) that the system of curves defined by

$$\frac{dx^\alpha}{dt} = g^{\alpha\beta}K_\beta \tag{9}$$

has an invariant determination in the two-dimensional projective space, i.e., *the congruence given by* (9) *is independent of the particular metric tensor by which the geodesics of the space are determined.* The family of curves (9) is of especial interest since in general these curves are distinct from the invariant geodesics of the space. It would appear that this invariant family of curves (or the above projective vector invariant) would be of particular significance in the further study of the geometry of the two-dimensional projective Riemann space.

[1] Weyl, H., "Zur Infinitesimalgeometrie: Einordnung der projektiven und der konformen Auffassung," *Göttingen Nachrichten*, 1921, pp. 99–112. A very complicated projective tensor of no stated significance has been constructed by J. M. Thomas, these PROCEEDINGS, **11**, 207–209 (1925).

[2] See, for example, Eisenhart, L. P., *Riemannian Geometry*, Princeton University Press, 1926, pp. 131–135.

ON CONVERGENCE IN LENGTH

BY MIRIAM C. AYER

THE OHIO STATE UNIVERSITY AND WELLESLEY COLLEGE

Communicated June 13, 1945

1. The topics discussed in this note have their origin in the theory of arc length. We shall be concerned with triples of equations of the form $x = x(u), y = y(u), z = z(u)$ where each function is defined and continuous on a closed linear interval $\alpha \leq u \leq \beta$. For conciseness we write the triple as $\mathfrak{x} = \mathfrak{x}(u)$, $u \in I$, where $\mathfrak{x}(u)$ is the vector $(x(u), y(u), z(u))$ and I is the interval $\alpha \leq u \leq \beta$. With length defined in the usual way in terms of inscribed polygons, the length of the curve represented by $\mathfrak{x} = \mathfrak{x}(u)$, $u \in I$, is then a function $L(\mathfrak{x})$ of the vector $\mathfrak{x}$.

The following definitions explain the notation and terminology used. The magnitude of a vector $\mathfrak{x}$ is denoted by $|\mathfrak{x}|$. Given any vector function $\mathfrak{x}(u) = (x(u), y(u), z(u))$, $u \in I$, the vector $(x'(u), y'(u), z'(u))$, defined wherever these derivatives exist, is written as $\mathfrak{x}'(u)$ or, more briefly, $\mathfrak{x}'$. A vector $\mathfrak{x}(u)$, $u \in I$, is said to be continuous if and only if each component $x(u)$, $y(u)$, $z(u)$ is continuous on I. A vector $\mathfrak{x}(u)$, $u \in I$, is said to be of bounded variation (briefly, $\mathfrak{x}$ is BV on I) if and only if each component

$x(u)$, $y(u)$, $z(u)$ is of bounded variation on I. A vector $\mathfrak{x}(u)$, $u \epsilon I$, is said to be absolutely continuous (briefly, AC) if and only if each component $x(u)$, $y(u)$, $z(u)$ is absolutely continuous on I. The functional $L(\mathfrak{x})$ is defined explicitly as follows: $L(\mathfrak{x}) = lub\Sigma|\mathfrak{x}(u'') - \mathfrak{x}(u')|$, where the least upper bound is taken over all finite subdivisions of I into non-overlapping intervals $u' \leq u \leq u''$. For vectors of the special form $\mathfrak{x}(u) = (f(u), 0, 0)$ the length reduces to the total variation $V(f)$ of $f(u)$ on I.

We state here some well-known facts about arc length.[6] $L(\mathfrak{x})$ is finite if and only if $\mathfrak{x}$ is BV on I. If $L(\mathfrak{x})$ is finite, then $\mathfrak{x}'(u)$ exists a.e. (almost everywhere) on I, $|\mathfrak{x}'(u)|$ is summable on I, and $L(\mathfrak{x}) \geq \int_I |\mathfrak{x}'|$. Equality holds in this relation if and only if $\mathfrak{x}$ is AC on I.

Uniform convergence on I of a sequence $f_n(u)$, $u \epsilon I$, $n = 0, 1, 2, \ldots$ is written as $f_n \to f_0\{\mathfrak{U}\}$. Similarly, $f_n \to f_0\{\mathfrak{M}\}$ on E denotes convergence in measure of f_n to f_0 on the set E. A sequence of vectors $\mathfrak{x}_n(u) = (x_n(u)$, $y_n(u)$, $z_n(u))$, $u \epsilon I$, $n = 0, 1, 2, \ldots$ is said to converge uniformly (briefly, $\mathfrak{x}_n \to \mathfrak{x}_0 \{\mathfrak{U}\}$) if and only if $x_n(u)$, $y_n(u)$, and $z_n(u)$ converge uniformly on I to $x_0(u)$, $y_0(u)$ and $z_0(u)$, respectively. It is well known[1, 2, 5] that if $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{U}\}$, $\mathfrak{x}_n$ BV and continuous for $n = 0, 1, 2, \ldots$, then $\liminf L(\mathfrak{x}_n) \geq L(\mathfrak{x}_0)$. When $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{U}\}$ and $L(\mathfrak{x}_n) \to L(\mathfrak{x}_0)$, then we shall say that $\mathfrak{x}_n$ converges to $\mathfrak{x}_0$ in length, in symbols: $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{L}\}$. Hereafter all vectors $\mathfrak{x}(u)$, $u \epsilon I$, will be assumed to be both BV and continuous on I.

In the literature on this subject[2, 5] previous results relate convergence in length with the following: (i) convergence in direction, i.e., convergence in some appropriate sense of $\mathfrak{x}_n'(u)$ to $\mathfrak{x}_0'(u)$, and (ii) some kind of convergence of each of the sequences $x_n(u)$, $y_n(u)$ and $z_n(u)$, $n = 0, 1, 2, \ldots$. In these results there is a sharp distinction between the parametric and the non-parametric cases, the latter being the special case in which $\mathfrak{x}_n(u) = (u$, $y_n(u)$, $z_n(u))$ for $n = 0, 1, 2, \ldots$. The non-parametric case appears to be somewhat simpler than the more general parametric case and certain results valid in the non-parametric case seem to admit of no extension to the parametric case. The purpose of this paper is to make further contributions to the theory along these lines and, in particular, to narrow the gap between the parametric and non-parametric cases.

2. In previous literature[2, 5] convergence in variation and strong convergence in variation are considered along with convergence in length. For sequences of vectors, however, we shall use strong convergence in length instead of strong convergence in variation. (These two conceptions are equivalent, and for vectors of the special form $\mathfrak{x}_n = (f_n(u), 0, 0)$ strong convergence in length of $\mathfrak{x}_n$ reduces to strong convergence in variation of the scalars $f_n(u)$.) We state the definitions of the types of convergence used in this paper. A sequence $f_n(u)$, $u \epsilon I$, f_n BV and continuous on I, $n = 0, 1, 2, \ldots$, is said to converge in variation, briefly $f_n \to f_0\{\mathfrak{V}\}$, if and only if $f_n \to f_0\{\mathfrak{U}\}$ and $V(f_n) \to V(f_0)$. A sequence of vectors $\mathfrak{x}_n(u) =$

$(x_n(u), y_n(u), z_n(u))$, $n = 0, 1, 2, \ldots$, is said to converge in variation, briefly $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{V}\}$, if and only if $x_n \to x_0\{\mathfrak{V}\}$, $y_n \to y_0\{\mathfrak{V}\}$, and $z_n \to z_0\{\mathfrak{V}\}$. A sequence of vectors $\mathfrak{x}_n(u)$, $\mathfrak{x}_n$ BV and continuous on I, $n = 0, 1, 2, \ldots$, is said to converge strongly in length, briefly $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{SL}\}$, if and only if $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{V}\}$ and $L(\mathfrak{x}_n - \mathfrak{x}_0) \to 0$.

It is easily verified that strong convergence in length implies convergence in length, but not conversely. Moreover it is well known[2,5] that convergence in variation is necessary but not sufficient for convergence in length. Results stated in later sections will clarify further the relations among these convergence types.

3. If $\mathfrak{x}(u)$ is any given vector continuous and BV on I, each subinterval $\Delta \subset I$ determines an arc of length $L(\Delta, \mathfrak{x})$. In this notation the quantity previously denoted by $L(\mathfrak{x})$ is written as $L(I, \mathfrak{x})$. The interval function $L(\Delta, \mathfrak{x})$ is non-negative, continuous and additive, and therefore can be extended to a c.a. (completely additive) function $L(E, \mathfrak{x})$ of Borel sets E in I. By the general theory of such set functions we then have the Lebesgue decomposition[6] $L(E, \mathfrak{x}) = L_a(E, \mathfrak{x}) + L_s(E, \mathfrak{x})$, where $L_a(E, \mathfrak{x}) = \int_E |\mathfrak{x}'|$ is a non-negative, c.a., AC function of Borel sets and $L_s(E, \mathfrak{x})$ is a non-negative, c.a., singular function of Borel sets. A BV function $f(u)$, $u \in I$, is expressible as the sum of an AC function and a singular function, this decomposition being univocally determined if we agree that the singular part vanish at the left end point of I. This unique decomposition will be called the normal Lebesgue decomposition of f and will be denoted by $f = f_a + f_s$, where f_a is AC and f_s singular. In terms of the normal Lebesgue decompositions of the components $x(u)$, $y(u)$, $z(u)$ we then introduce the normal Lebesgue decomposition of the vector $\mathfrak{x}$ as $\mathfrak{x}_a + \mathfrak{x}_s$, where $\mathfrak{x}_a = (x_a, y_a, z_a)$ and $\mathfrak{x}_s = (x_s, y_s, z_s)$. The vectors $\mathfrak{x}_a$ and $\mathfrak{x}_s$ give rise, in turn, to non-negative, c.a. functions of Borel sets $L(E, \mathfrak{x}_a)$ and $L(E, \mathfrak{x}_s)$, respectively. The two Lebesgue decompositions $L(E, \mathfrak{x}) = L_a(E, \mathfrak{x}) + L_s(E, \mathfrak{x})$ and $\mathfrak{x}(u) = \mathfrak{x}_a(u) + \mathfrak{x}_s(u)$ are related in the following simple way.

LEMMA: *If $\mathfrak{x}(u)$ is BV and continuous on I, then $L_a(E, \mathfrak{x}) = L(E, \mathfrak{x}_a)$ and $L_s(E, \mathfrak{x}) = L(E, \mathfrak{x}_s)$ for every Borel set $E \subset I$.*

The Lemma of the Lebesgue decomposition leads to a number of new results and makes it possible to simplify proofs of several known results. By way of illustration we state some of the implications of the lemma.

(a) If $\mathfrak{x}_n(u)$, $u \in I$, $n = 0, 1, 2, \ldots$, is a sequence of continuous, BV vectors, then $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{SL}\}$ if and only if $\mathfrak{x}_{na} \to \mathfrak{x}_{0a}\{\mathfrak{SL}\}$ and $\mathfrak{x}_{ns} \to \mathfrak{x}_{0s}\{\mathfrak{SL}\}$.

(b) If $\mathfrak{x}_1(u)$ and $\mathfrak{x}_2(u)$ are BV and continuous on I, then $L(\mathfrak{x}_1 + \mathfrak{x}_2) = L(\mathfrak{x}_1) + L(\mathfrak{x}_2)$ if and only if $L(\mathfrak{x}_{1a} + \mathfrak{x}_{2a}) = L(\mathfrak{x}_{1a}) + L(\mathfrak{x}_{2a})$ and $L(\mathfrak{x}_{1s} + \mathfrak{x}_{2s}) = L(\mathfrak{x}_{1s}) + L(\mathfrak{x}_{2s})$.

(c) If $\mathfrak{x} = (u, f(u), 0)$, $f(u)$ continuous and BV on I, then $L(\mathfrak{x}) \leq (\beta - \alpha) + V(f)$, equality holding if and only if $f(u)$ is singular.

(*d*) If $\mathfrak{x}_n$ is BV and continuous for $n = 1, 2, \ldots$ and $\mathfrak{x}_0$ is AC, then $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{SL}\}$ if and only if $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{U}\}$ and $|\mathfrak{x}_n' - \mathfrak{x}_0'| \to 0\{\mathfrak{M}\}$ on I.

(*e*) If $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{U}\}$, $\mathfrak{x}_n$ AC for $n = 1, 2, \ldots$, then $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{SL}\}$ if and only if $\int_I |\mathfrak{x}_n' - \mathfrak{x}_0'| \to 0$.

(*f*) If $\mathfrak{x} = \mathfrak{x}_n(u)$, $n = 1, 2, \ldots$, represent polygons inscribed in the curve represented by $\mathfrak{x} = \mathfrak{x}_0(u)$, where $\mathfrak{x}_0$ is BV and continuous, then $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{SL}\}$ if and only if $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{U}\}$ and $\mathfrak{x}_0$ is AC.

Let us add a remark concerning statement (*c*). In the special case where $f(u)$ is monotone and singular, (*c*) implies that $L(\mathfrak{x}) = (\beta - \alpha) + |f(\beta) - f(\alpha)|$. That is, the length of the curve $y = f(x)$ is then equal to the sum of its projections on the x and y axes. This can be regarded as a generalization of the well-known fact[2] that the curve determined by the Cantor function $y = f(x)$, $0 \leq x \leq 1$, has length 2.

4. The Steiner inequality is an important tool in previous literature.[3, 4] This inequality states that $L([\mathfrak{x}_1 + \mathfrak{x}_2]/2) \leq [L(\mathfrak{x}_1) + L(\mathfrak{x}_2)]/2$ for any two vectors $\mathfrak{x}_1(u)$, $\mathfrak{x}_2(u)$ which are BV and continuous on I. Although the denominator 2 is essential if we confine our attention to vectors in the non-parametric form, it is generally more convenient to omit it.

It is of interest to find conditions under which the sign of equality will hold. (See §2 (*b*).) In the non-parametric case we have the well-known simple result: If $\mathfrak{x}_1 = (u, f_1(u), 0)$, $\mathfrak{x}_2 = (u, f_2(u), 0)$, where $f_1(u)$ and $f_2(u)$ have continuous derivatives on I, then equality holds in the Steiner inequality if and only if $f_1(u)$ and $f_2(u)$ differ by a constant. That is, for equality to hold two such curves must be parallel. The following statement is a generalization of this.

(*a*) If $\mathfrak{x}_1(u)$ and $\mathfrak{x}_2(u)$ are AC on I, then $L(\mathfrak{x}_1 + \mathfrak{x}_2) = L(\mathfrak{x}_1) + L(\mathfrak{x}_2)$ if and only if $\mathfrak{x}_1'/|\mathfrak{x}_1'| = \mathfrak{x}_2'/|\mathfrak{x}_2'|$ a.e.$_2$ on $E = \underset{u}{E}[u \in I, |\mathfrak{x}_1'| > 0, |\mathfrak{x}_2'| > 0]$.

In the literature[3, 4] there are various generalizations involving a study of finding conditions under which the sign of equality will hold approximately. That work is concerned solely with the non-parametric case. The following statements may be considered as extensions of those results to the general case.

(*b*) If $\mathfrak{x}_1(u)$ and $\mathfrak{x}_2(u)$ are BV and continuous on I, then each of the following hold:

(i) $[\int_I \{|\mathfrak{x}_1'|\cdot|\mathfrak{x}_2'| - \mathfrak{x}_1'\mathfrak{x}_2'\}^{1/2}]^2 \leq [L(\mathfrak{x}_1) + L(\mathfrak{x}_2)]\cdot[L(\mathfrak{x}_1) + L(\mathfrak{x}_2) - L(\mathfrak{x}_1+\mathfrak{x}_2)]$

(ii) $[\int_I \{(\mathfrak{x}_1')^2(\mathfrak{x}_2')^2 - (\mathfrak{x}_1'\mathfrak{x}_2')^2\}^{1/4}]^4 \leq {}^1/_2[L(\mathfrak{x}_1) + L(\mathfrak{x}_2)]^3\cdot[L(\mathfrak{x}_1) + L(\mathfrak{x}_2) - L(\mathfrak{x}_1 + \mathfrak{x}_2)]$

In the non-parametric case where $\mathfrak{x}_1 = (u, y_1(u), z_1(u))$ and $\mathfrak{x}_2 = (u, y_2(u), z_2(u))$ we have $\mathfrak{x}_1' = (1, y_1'(u), z_1'(u))$ and $\mathfrak{x}_2' = (1, y_2'(u), z_2'(u))$. In this case an elementary discussion yields the relation $|\mathfrak{x}_1' - \mathfrak{x}_1'|^2 \leq (\mathfrak{x}_1')^2(\mathfrak{x}_2')^2 - (\mathfrak{x}_1'\mathfrak{x}_2')^2$. From part (ii) of (*b*) we then obtain the following result.

(*c*) If $\mathfrak{x}_1 = (u, y_1(u), z_1(u))$ and $\mathfrak{x}_2 = (u, y_2(u), z_2(u)$ are BV and continuous on I, then $\{\int_I |\mathfrak{x}_1' - \mathfrak{x}_2'|^{1/2}\}^4 \leq \{L(\mathfrak{x}_1) + L(\mathfrak{x}_2)\}^3\cdot\{[L(\mathfrak{x}_1) + L(\mathfrak{x}_2)]/2 -$

$L([\mathfrak{x}_1 + \mathfrak{x}_2]/2)\}$. This inequality is the analogue of a known inequality[4] involving surfaces. Results known for the non-parametric case now appear as corollaries of new results for the parametric case.

The following statements are corollaries of inequalities (b) and (c).

(d) If $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{L}\}$, then $|\mathfrak{x}_n'|.|\mathfrak{x}_0'| - \mathfrak{x}_n'\mathfrak{x}_0' \to 0\{\mathfrak{M}\}$ on I and $(\mathfrak{x}_n')^2(\mathfrak{x}_0')^2 - (\mathfrak{x}_n'\mathfrak{x}_0')^2 \to 0\{\mathfrak{M}\}$ on I.

(e) If $\mathfrak{x}_n = (u, y_n(u), z_n(u))$ for $n = 0, 1, 2, \ldots$ and if $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{L}\}$, then $|\mathfrak{x}_n' - \mathfrak{x}_0'| \to 0\{\mathfrak{M}\}$ on I.

5. Statement (e) of §4 is an example of a case where convergence in length implies some kind of convergence of the derivatives. That statement is concerned with the non-parametric case and we would like to find its analogue for the parametric case. Let t_n denote the unit tangent vector $\mathfrak{x}_n'/|\mathfrak{x}_n'|$, defined wherever $|\mathfrak{x}_n'| > 0$, $n = 0, 1, 2, \ldots$. Results concerning the behavior of t_n are obtained from (d) of § 4.

(a) If $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{L}\}$, then $\int_E |t_n - t_0| \to 0$, where $E = \underset{u}{E}[u \in I, |\mathfrak{x}_0'| > 0, \liminf |\mathfrak{x}_n'| > 0]$.

(b) If $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{L}\}$ and $\mathfrak{x}_n = (u, y_n(u), z_n(u))$ for $n = 0, 1, 2, \ldots$, then $\int_I |t_n - t_0| \to 0$.

In a sense, the preceding statements can be considered as relating convergence in length to convergence in direction. Although the converses are not true, the following special case is of interest.

(c) If $\mathfrak{x}_n = (u, y_n(u), z_n(u))$, $u \in I$, $n = 0, 1, 2, \ldots$, and $\mathfrak{x}_0$ is AC, then $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{L}\}$ if and only if $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{U}\}$ and $|\mathfrak{x}_n' - \mathfrak{x}_0'| \to 0$ $\{\mathfrak{M}\}$ on I.

(d) If $\mathfrak{x}_n = (u, y_n(u), z_n(u))$, $u \in I$, $n = 0, 1, 2, \ldots$, and $\mathfrak{x}_0$ is AC, then $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{L}\}$ if and only if $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{U}\}$ and $\int_I |t_n - t_0| \to 0$.

6. There is apparently a lack of finality and completeness in the theory of convergence in length. The reason for this is that some of the simple properties of strong convergence in length do not generally hold for convergence in length. In special cases where suitable additional conditions are imposed on the vector functions complete characterizations are possible. Statements (c) and (d) of § 5 illustrate such a case.

A sequence $\mathfrak{x}_n(u) = (x_n(u), y_n(u), z_n(u))$, $u \in I$, $n = 0, 1, 2, \ldots$ is said to be uniformly AC if and only if $x_n(u)$, $y_n(u)$, $z_n(u)$, $n = 0, 1, 2, \ldots$ are equi-absolutely continuous on I. In the non-parametric case this is a necessary condition for a sequence of AC vectors to converge in length.

(a) If $\mathfrak{x}_n(u) = (u, y_n(u), z_n(u))$, $u \in I$, $\mathfrak{x}_n$ AC, $n = 0, 1, 2, \ldots$, then $\mathfrak{x}_n \to \mathfrak{x}_0\{\mathfrak{L}\}$ if and only if the vectors $\mathfrak{x}_n$, $n = 0, 1, 2, \ldots$ are uniformly AC and $|\mathfrak{x}_n' - \mathfrak{x}_0'| \to 0\{\mathfrak{M}\}$ on I.

In the non-parametric case the absolute continuity of $\mathfrak{x}_0$ is sufficient to make convergence in length equivalent to strong convergence in length. Obtaining characterization theorems for the cases just mentioned is therefore a simple problem. In the following result, however, strong convergence in length is not implied.

(*b*) If $\mathfrak{x}_n \rightarrow \mathfrak{x}_0\{\mathfrak{V}\}$, $\mathfrak{x}_n$ uniformly AC on I for $n = 0, 1, 2, \ldots$, then $\mathfrak{x}_n \rightarrow \mathfrak{x}_0\{\mathfrak{T}\}$ if and only if

$$\text{(i)}\quad |\mathfrak{x}_n'|\cdot|\mathfrak{x}_0'| - \mathfrak{x}_n'\mathfrak{x}_0' \rightarrow 0\{\mathfrak{M}\} \text{ on } I$$
$$\text{(ii)}\quad |\mathfrak{x}_n'| \rightarrow 0\{\mathfrak{M}\} \text{ on } E = \underset{u}{E}[u \in I, |\mathfrak{x}_0'| = 0].$$

7. For vectors of the special form $\mathfrak{x}_n = (f_n(u), 0, 0)$ the results of §§ 3–6 yield as corollaries corresponding statements on convergence in variation. We state a few of these.

(*a*) If $f_1(u)$ and $f_2(u)$ are BV and continuous on I, then $V(f_1 + f_2) = V(f_1) + V(f_2)$ if and only if $f_1'f_2' \geq 0$ a.e. on I and $V(f_{1s} + f_{2s}) = V(f_{1s}) + V(f_{2s})$.

(*b*) If $f_n \rightarrow f_0\{\mathfrak{V}\}$, then $|f_nf_0'| - f_nf_0' \rightarrow 0\{\mathfrak{M}\}$ on I.

(*c*) If $f_n \rightarrow f_0\{\mathfrak{V}\}$, f_n equi-absolutely continuous on I for $n = 0, 1, 2, \ldots$, then $f_n \rightarrow f_0\{\mathfrak{V}\}$ if and only if

$$\text{(i)}\quad |f_n'f_0'| - f_n'f_0' \rightarrow 0\{\mathfrak{M}\} \text{ on } I$$
$$\text{(ii)}\quad f_n' \rightarrow 0\{\mathfrak{M}\} \text{ on } E = \underset{u}{E}[u \in I, f_0' = 0].$$

[1] Adams, C. R., and Clarkson, J. A., *Bull. Am. Math. Soc.*, **40**, 413–417 (1934).
[2] Adams, C. R., and Lewy, H., *Duke Math. Jour.*, **1**, no. 1, 19–26 (1935).
[3] McShane, E. J., *Ann. Math.*, (2) **33**, 125–138 (1932).
[4] Radó, T., and Reichelderfer, P., *Bull. Am. Math. Soc.*, **47**, no. 2, 102–108 (1941).
[5] Radó, T., and Reichelderfer, P., *Duke Math. Jour.*, **9**, no. 3, 527–565 (1942).
[6] Saks, S., *Theory of the Integral*, Warszawa-Lwów, 1937.

PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 31 September 15, 1945 Number 9

*PURE STRAIN MICE BORN TO HYBRID MOTHERS FOLLOWING OVARIAN TRANSPLANTATION**

BY W. L. RUSSELL AND JANE GOODRICH HURST†

ROSCOE B. JACKSON MEMORIAL LABORATORY, BAR HARBOR, MAINE

Communicated July 20, 1945

It is well established that many characters in mammals are affected by naturally occurring variation in the uterus or, to use a more inclusive term, the prenatal maternal environment. Although this factor is regarded as environmental as far as the embryo is concerned, it is itself presumably subject to both environmental and genetic factors (Russell[1]).

The work of Wright,[2] Wright and Chase,[3] Green [4] and others,[5] who made statistical analyses of variation within inbred strains of mammals, clearly shows that the characters which they studied, polydactyly, white spotting, number of vertebrae, etc., were affected by environmentally determined variation in the prenatal maternal environment.

The effect of genetically determined variation in the prenatal maternal environment on character variation has, however, received little attention, perhaps because of a lack of easy methods of evaluating it. Evidence that this factor may be important is provided by the increasing number of cases in which a difference has been observed between reciprocal F_1 hybrids obtained by crossing inbred strains. Russell and Green[6] have reported one such case for number of lumbar vertebrae in mice. Similar data from other crosses have been obtained and are in preparation for publication. The mere discovery of such a difference does not, however, furnish proof of a difference between the maternal environments of the parental strains: without further analysis the difference between the hybrids could equally well be attributed to the egg cytoplasm or, in the heterogametic sex, to sex linked genes.

Two methods have been developed which can be used to test directly for differences between prenatal maternal environments. The first has so far been utilized for this purpose on a large scale only by Fekete and Little.[7] In their work, fertilized ova are removed from the oviduct of one inbred

strain of mice and transferred, by means of a fine glass pipette, to the uterus of another inbred strain. The resulting offspring thus develop from the chromosomes and cytoplasm of the first strain and in the uterine environment of the second.

The second method, developed recently by Robertson,[8] makes ingenious use of transplantation of ovaries within an inbred strain maintained by forced heterozygosis. In Robertson's experiments, ovaries from yellow (A^yA^w) mice were transplanted to agouti (A^wA^w) mice of the same inbred strain. By mating these, and unoperated yellow females, with yellow males he was able to compare the development of the lethal homozygous A^yA^y type in the uterus of an agouti with its development in the uterus of a yellow.

The two methods differ in that the transplantation of ovaries brings the foster mother's environment into play at an earlier stage. Whether this is important or not is not known, but a combination of the two methods on the same material could be used to investigate this point. The methods differ also in their practical limitations. The small average number of offspring obtained from operations with transferred ova would make the collection of the large number of individuals often required for the study of quantitative variation somewhat tedious. Ovarian transplantation, while more promising in this respect, seemed to be limited to studies of single factor genetic differences within inbred strains maintained by forced heterozygosis. It was presumably not available for the investigation of multiple factor differences between strains because of the failure of tissue grafts in foreign hosts. That interstrain transplants of mouse ovaries are unsuccessful has been shown by Robertson[8] in a total of 36 operations and by Mr. Ralph Kellogg† at this laboratory in a total of 23 operations.

An attempt to find a method by which ovarian transplantation could be used to measure the effect of multiple genetic differences in the maternal environment led to the experiments described in this paper. It occurred to us that if ovaries could be transplanted from an inbred strain to the F_1 hybrid of that and another strain, a type of transplant that is usually successful, this would provide a method for comparing pure strain and hybrid maternal environments.

For the first experiment, strain dba ovaries were transplanted to dba/C3H F_1 hybrids which were then mated with dba males. These strains were chosen because the dba strain (genotype *ddbbaa*) differs by three recessive color factors from the C3H strain (*DDBBAA*), a fact which, as shown in figure 1, would help in distinguishing successful operations from those cases in which the hybrid host ovary regenerated. Thus offspring from the transplanted ovary would all be dilute brown non-agouti (*ddbbaa*), while only one-eighth of the offspring from a host's regenerated ovary would be of this color. In all operations, except those on females Nos. 162

and 163 (see table 1), we followed the essentials of the technique described by Robertson.[8] The left ovary, capsule and oviduct of each host were removed. Then, 7 to 15 days later, a second operation was performed in which the host's right ovary was removed from its capsule and the donor ovary implanted in its place. In female No. 162 the first operation was omitted, one host ovary being left intact, and in No. 163 there was no interval between the two operations. The latter variation in technique is discussed later in this paper.

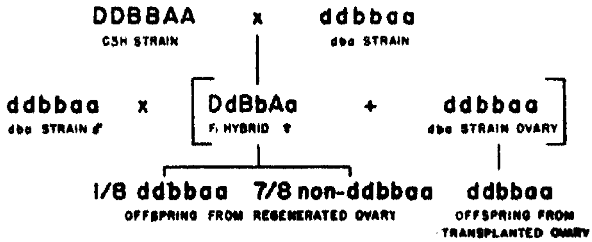


FIGURE 1

Of the 22 operated hybrid females, 9 had no offspring and one produced a single litter which was eaten shortly after birth. The remaining 12 gave the results shown in table 1. Females Nos. 17, 23 and 58 clearly show re-

TABLE 1

NUMBER AND TYPE OF OFFSPRING FROM MATINGS OF dba MALES WITH dba/C3H F_1 HYBRID FEMALES INTO WHICH dba OVARIES HAD BEEN IMPLANTED

Serial No. of hybrid host	17	23	25	26	45	55	58	65	162	163	193	197
ddbbaa offspring	..	..	2	2	19	4	..	18	1	12	9	8
Non-*ddbbaa* offspring	6	3	8	17	..	24	7	4	11	..	...	..

generation of the host ovary and failure of the transplanted ovary. Females Nos. 25, 26, 55 and 162 also indicate failure of the graft, for the proportion of *ddbbaa* offspring produced by each of these females is not significantly different from one-eighth. In females Nos. 45, 163, 193 and 197 the transplanted ovary was clearly functional and there was apparently no regeneration of the host ovary. The remaining female, No. 65, produced 4 non-*ddbbaa* offspring which must have come from a regenerated host ovary, but the high proportion of *ddbbaa* offspring, 18 out of 22, indicates that the transplanted ovary was also functioning.

Robertson[8] does not mention the possibility of regeneration occurring along with a functional graft in his results, but his ratio of 38 yellow to 62 agouti, where a 1:1 ratio was expected from the transplanted ovary, indicates that regeneration probably occurred in some of his females and a ratio of 38 yellow to 52 agouti from those females which produced at least one yellow offspring, though not significantly different from a 1:1 ratio,

does differ in the direction that favors the possibility of regeneration having occurred even in the presence of a functional graft.

In our work, in any case, subsequent experiments have shown that this type of result is not uncommon and apparently cannot be avoided even by careful operative technique. This is unfortunate because in the above experiments, for example, even where the success of the transplant was established by the fact that only *ddbbaa* animals were born, one cannot be certain that any particular *ddbbaa* offspring was not produced from a regenerated host ovary. With hybrids from strains differing by a smaller number of marker genes there would be even greater uncertainty, as we found in a series of operations made by the junior author and Miss Barbara Perry.† In this series C57 black (genotype *aa*) mice were crossed with C3H agouti (*AA*). C57 black ovaries were transplanted to the F_1 hybrids which were then mated with C57 black males. The expected ratio in the offspring from a regnerated hybrid ovary is thus one half agouti (*Aa*) and one half black (*aa*), while the transplanted ovary would yield only black offspring. Of the 23 animals that had young, only one gave reasonably clear-cut results. This produced 3 agouti and 18 black offspring, indicating a combination of successful transplant and regenerated host ovary. Of the remaining 22, 8 gave only black offspring (a total of 26) and 14 gave both agouti and black (totaling 56 and 70, respectively). The total of 56 agouti to 96 black produced by these 22 females indicates that the ovarian grafts were functional in at least some, but the number of offspring was not large enough to be sure of this in any particular case.

It is apparent that the ever-present possibility of regeneration of the host ovary occurring along with a successful graft is a serious limitation. The difficulty, of course, resides in the fact that the hybrid carries the marker genes of both parental strains and will segregate each parental combination in at least some of its germ cells. To get around this difficulty a method was worked out which makes use of an inbred strain of mice which we have called "stock 129" and which was originally obtained from Prof. L. C. Dunn. This strain was inbred with forced heterozygosis to carry both chinchilla (c^{ch}) and albino (c^a) alleles. It also carries pink-eye and light-bellied agouti genes, but this can be disregarded in the present discussion. It is now maintained, as shown in figure 2 (*a*), by brother-sister matings of the heterozygotes ($c^{ch}c^a \times c^{ch}c^a$), thus producing, in addition to the heterozygotes, the two homozygous combinations $c^{ch}c^{ch}$ and c^ac^a. All three types are phenotypically distinguishable. Two ways in which this strain is being used in ovarian transplant experiments are shown in figure 2. In figure 2 (*b*) the F_1 hybrid is obtained by mating any full colored, *CC*, strain with the stock 129 $c^{ch}c^{ch}$ type. Ovaries from stock 129 c^ac^a animals are transplanted to the Cc^{ch} hybrids which are then mated with stock 129 $c^{ch}c^{ch}$ males. Offspring from a regnerated host ovary will thus be either Cc^{ch} or $c^{ch}c^{ch}$, genotypes

which are phenotypically distinguishable from the $c^{ch}c^{a}$ combination produced by a successful ovarian graft. The origin of every individual born to the hybrid can, therefore, be determined. Figure 2 (c) shows a similar method that can be applied to a cross with any albino strain. The system of matings with the strains shown in figure 2 can be worked out in other ways, but the two examples given serve to illustrate the principle.

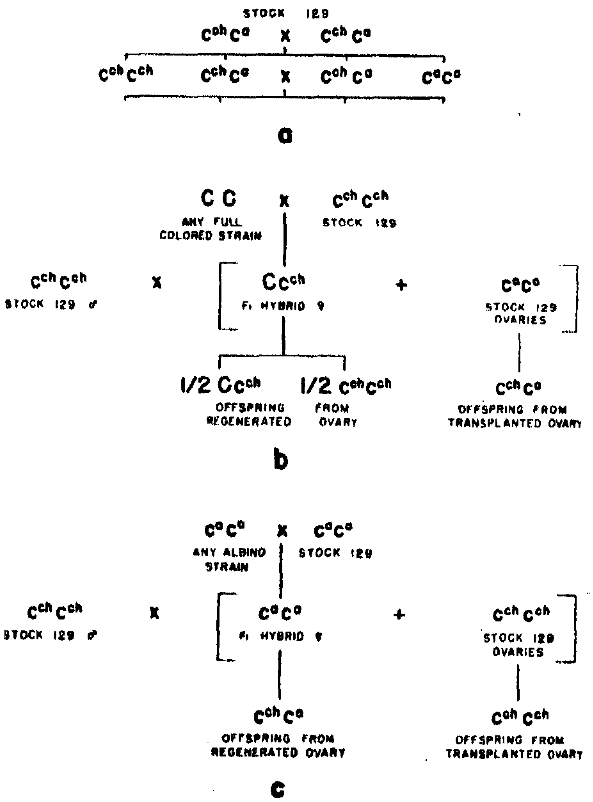


FIGURE 2

(a) System of maintaining stock 129. (b, c) Two ways in which this stock can be used in ovarian transplantations to hybrid females so that offspring of the grafted ovary can be distinguished from offspring of a regenerated host ovary.

There are other possibilities. Thus $c^{ch}c^{ch}$ or $c^{e}c^{e}$ strains could be used just as well as the CC and $c^{a}c^{a}$ strains, and strains carrying other genes maintained in forced heterozygosis could be used in place of stock 129. As a control for possible effects of the operation itself, transplants can be

made both ways between the two homozygous types of the 129 stock or whatever stock is used in its place.

In the above systems, pure strain mice are raised in hybrid hosts by taking the transplanted ovary, as well as the male with which its host is mated, from the same strain. It is, however, equally possible to use males from other strains, thus producing hybrid offspring raised in hybrid uteri. One case of this is reported below. Of course the choice of strains is again restricted to those carrying suitable marker genes.

Summarizing the above, it is clear that, within the limitation of having to choose appropriate strains of animals, a method is provided by which ovarian transplantation can be used for comparing pure strain and hybrid maternal environments. The method could also be used in other fields of biology wherever, for example, an experiment requires a means of distinguishing between the functioning of host and graft ovarian tissue.

In our experiments, ovaries from stock 129 have been transplanted to hybrids of this stock and four other stocks: C3H, dba, C57 black and Bagg albino. The first three are full colored (*CC*) strains and were used as shown in figure 2 (*b*), while the Bagg albino was mated as in figure 2 (*c*). Functional grafts have been obtained in all of these hybrids. In most cases the hybrids were mated to stock 129 males, thus yielding pure stock 129 animals raised in four different hybrid maternal environments. Some of the 129/C57 hybrids were mated with Bagg albino males to produce, from the transplanted ovary, 129/Bagg hybrids developed in 129/C57 uteri.

Details of technique, and findings in regard to the effect of various factors on the success of the transplantations, will be presented in another paper, but it seems desirable to mention here the following departure from Robertson's technique. He removed one host ovary 7 to 10 days before the transplantation in order to produce hypertrophy of the remaining ovary and of its capsule into which the ovary of the donor was to be placed, but he questions whether this is necessary and suggests testing it by removal of both host ovaries at the time of transplantation. The success obtained with female No. 163 (table 1), which was operated upon in this way, shows that hypertrophy of the site of implantation is not an essential factor. In view of this finding, most of the experiments with stock 129 have been made by removing both ovaries of a host and implanting donor ovaries in their place in the same operation. In our work, operations have been more successful with this than with Robertson's method, possibly because two graft ovaries per operated female are used instead of one.

At the present time about 40% of over 100 operations made by the senior author on stock 129 hybrids have resulted in successful grafts. Nearly one-half of these have shown at least some regeneration of host ovarian tissue. The number of offspring obtained from the grafted ovaries has been very high in some cases. The highest number recorded for a single female is 80

This was a 129/Bagg hybrid containing stock 129 ovaries and mated with a stock 129 male. The number of young is higher than has ever been recorded at this laboratory for a straight stock 129 mating. (The operation was made by Miss Winona Hinkley† who assisted in the early work with the 129 stock.)

A large series of experiments is being conducted to compare the maternal environments of the 129 stock and some of its hybrids, using number of vertebrae as a character. It has already been stated that differences in number of lumbar vertebrae obtained in reciprocal hybrids of certain strains indicate that this character may be affected by the maternal environment.

Summary.—The only way, so far reported, of testing directly for the effect of multiple gene, or strain, differences in maternal environments is by the somewhat tedious process of transferring fertilized ova from one strain to another. The work reported here provides a method of using ovarian transplantation for this purpose, at least as far as a comparison of pure strain and hybrid maternal environments is concerned.

Offspring were obtained from pure strain ovaries transplanted to hybrid females. Six different host-donor combinations, involving a total of three inbred strains and six hybrids, were tried and all were successful. In many cases a successful graft was accompanied by a regenerated hybrid host ovary. In early experiments it was not possible to distinguish offspring of the grafted and regenerated ovaries except on a statistical basis, but mating plans were devised using strains with suitable marker genes in such a way that it is possible to determine the origin of all the offspring by their phenotype.

* Aided by a grant from the Rockefeller Foundation and by grants to the Roscoe B. Jackson Memorial Laboratory from the National Cancer Institute, the Jane Coffin Childs Memorial Fund for Medical Research, the Anna Fuller Fund and the International Cancer Research Foundation.

† Summer student at the Roscoe B. Jackson Memorial Laboratory.

[1] Russell, W. L., Chap. 10 in Staff of the Roscoe B. Jackson Memorial Laboratory, *Biology of the Laboratory Mouse*, Blakiston Co., Philadelphia, 497 pp. (1941).

[2] Wright, S., *Genetics*, 19, 506–536 (1934).

[3] Wright, S., and Chase, H. B., *Ibid.*, 21, 758–787 (1936).

[4] Green, E. L., *Ibid.*, 26, 192–222 (1941).

[5] See Green[4] for references.

[6] Russell, W. L., and Green, E. L., *Genetics*, 28, 87 (1943).

[7] Fekete, E., and Little, C. C., *Cancer Res.*, 2, 525–530 (1942).

[8] Robertson, G. G., *Jour. Exp. Zool.*, 89, 197–232 (1942).

EXPERIMENTS ON SEXUAL ISOLATION IN DROSOPHILA. V. THE EFFECT OF VARYING PROPORTIONS OF DROSOPHILA PSEUDOOBSCURA AND DROSOPHILA PERSIMILIS ON THE FREQUENCY OF INSEMINATION IN MIXED POPULATIONS

BY HOWARD LEVENE AND TH. DOBZHANSKY*

COLUMBIA UNIVERSITY

Communicated July 19, 1945

Introduction.—When females of *Drosophila pseudoobscura* and *Drosophila persimilis* are placed together with males of one of these species, it is usually found that a greater proportion of the conspecific than of the alien females are inseminated. The nature of the stimuli that lead to this result is still obscure, although Mayr and Dobzhansky[1] have shown that the degree of the preference varies depending upon the history of the individual flies involved and upon the environment. The experiments to be reported in the present article are concerned with the effects of the relative numbers of the conspecific and of the alien females available to the males on the frequencies of homogamic and heterogamic matings. Perhaps the simplest of the many hypotheses that may be constructed about the matings which take place in mixed populations is that when a female and a male meet, there is a certain definite probability that they will mate, depending on their respective species but not on what other males and females may be found in the same medium. On the other hand, it is also possible that the flies are stimulated by the proximity of conspecific individuals of the opposite sex in such a manner that they become more receptive to mating with alien as well as with conspecific partners than they would otherwise be. Still another possibility is that only homogamic matings take place while conspecific individuals of the opposite sex are available, and that heterogamic matings occur only, or predominantly, when the possibility of homogamic mating is excluded or remote. The above possibilities are not necessarily mutually exclusive, and various intermediate situations may be encountered. It is also fully conceivable that different pairs of species and races might behave quite differently in this respect.

Methods and Procedure.—The orange strain of *D. pseudoobscura* and the Stony Creek strain of *D. persimilis* were used in all experiments.[1] In the main body of seven experiments, 10 freshly hatched *D. persimilis* males were placed in each vial together with the following freshly hatched females: (1) 10 *persimilis*, (2) 10 *pseudoobscura*, (3) 10 *persimilis* and 10 *pseudoobscura*, (4) 20 *persimilis* and 10 *pseudoobscura*, (5) 40 *persimilis* and 10 *pseudoobscura*, (6) 10 *persimilis* and 20 *pseudoobscura*, and (7) 10 *persimilis* and 40 *pseudoobscura*. In other words, only *D. persimilis* males were used in all vials and their numbers were always ten per vial; the numbers and

the species of the females were varied. The vials were kept in an incubator at 25½°C. for approximately 4 days, whereupon the females were dissected and the presence or absence of sperm in their seminal receptacles was determined by microscopic examination. These experiments were done in September–October, 1944, with the assistance of Mr. George Streisinger, whose help we wish to acknowledge. The results are summarized as experiments 1 to 7 in table 1.

TABLE 1

Numbers of Inseminated Females Observed When 10 *D. persimilis* Males Were Placed with n_c *D persimilis* Females and n_a *D. pseudoobscura* Females

			D. persimilis			*D. pseudoobscura*				
Expt. Number x	No. of Conspecific per Vial n_{cx}	No. of Alien per Vial n_{ax}	Fertilized F_{cx}	Unfertilized $N_{cx}-F_{cx}$	Total N_{cx}	Fertilized F_{ax}	Unfertilized $N_{ax}-F_{ax}$	Total N_{ax}	Isolation Index b_x	Isolation Ratio g_x
			....*	....						
1	10	0	115*	28	143					
			80.41*	19.59						
						...	....			
2	0	10				28	125	153		
						18.30	81.70			
			143.67	34.33		27.34	145.66			
3	10	10	141	37	178	39	134	173	0.5569	0.2846
			79.21	20.79		22.54	77.46			
			149.68	41.33		14.27	78.73			
4	20	10	150	41	191	13	80	93	0.6978	0.1780
			78.53	21.47		13.98	86.02			
			204.67	168.33		9.67	80.33			
5	40	10	204	169	373	11	79	90	0.6347	0.2235
			54.69	45.31		12.22	87.78			
			57.03	44.97		22.99	187.01			
6	10	20	60	42	102	17	193	210	0.7581	0.1376
			58.82	41.18		8.10	91.90			
			87.72	41.28		70.17	456.83			
7	10	40	90	39	129	64	463	527	0.7035	0.1741
			69.77	30.23		12.14	87.86			
						...	....			
8	0	10				49	80	129		
						37.98	62.02			
			....	....		...	....			
9	10	10	122	6	128	67	63	130		
			95.31	4.69		51.54	48.46			

* For each experiment, the top line gives the "expected number" of flies on the basis of H2′ (see text), the middle line gives the observed numbers, and the bottom line gives the observed percentages.

Additional data were secured in April, 1945, with the aid of a slightly different method. Males and females of *D. persimilis* and *D. pseudoobscura* were aged in the absence of individuals of the opposite sex for a week to ten days, whereupon 10 *persimilis* males were placed in vials overnight with 10 *pseudoobscura* females or with 10 *pseudoobscura* and 10 *persimilis* females. (Experiments 8 and 9 in table 1.)

Similar experiments were performed on two geographic strains of *D. prosaltans*, but these will be discussed in a separate section.

When freshly hatched females and males were kept together for 3–4 days before dissection, the results obtained in different vials of the same experiment frequently showed statistically significant heterogeneity, while the experiments with the aged flies did not. The heterogeneity is, of course, a complicating factor in the analysis of the data. There are several probable sources of this heterogeneity. Since dissection and examination of the sperm receptacles is a rather laborious operation, females and males in some vials were kept together somewhat longer than in others. Some flies came from culture bottles in the early and others in late stages of hatching. Since the experiments of a given series extended for about two months, some variation in the food and other environmental factors may have occurred.

Analysis.—Throughout this discussion we will assume that there are 10 males of a given species or strain present in a standard vial, along with n_c conspecific[2] females and n_a alien females. Of these f_c and f_a, respectively, will be fertilized. Let N_c equal the sum of the n_c for a given experiment (i.e., N_c is the total number of conspecific females in all the vials of a given composition), and let N_a, F_c and F_a be the corresponding sums for n_a, f_c and f_a.

If π_c is the probability that any particular conspecific female will be inseminated during the course of the experiment, then π_c will be some function of n_c, n_a and of t, the time that the flies are left together. We denote this function by $\pi_c(n_c, n_a, t)$. We will include in t also such factors as temperature, age of the flies, food, etc., so that t represents what may be called the "physiological time." Similarly $\pi_a(n_c, n_a, t)$ is the probability that an alien female will be inseminated. Evidently π_c and π_a approach zero as the number of females per vial becomes large, since there must be some upper limit to the number of females a male can inseminate. Because of the variability from vial to vial of the per cent fertilized, as above discussed, the data do not give very precise information on the form of the functions π_c and π_a. However, from examination of the results for the individual vials it is clear that there is no very great variation in π_c over the range of values of n_c and n_a used, with the possible exception of a decrease in π_c in the experiment with 40 conspecific and 10 alien females. Apparently the decrease in percentage of fertilization does not become marked until considerably higher densities of females are reached than were used here. Since any analytic function is approximately linear for small changes in the variables, we may suppose that π_c depends on some linear combination of n_c and n_a when n_c and n_a do not vary too widely. We may represent this combination as $n' = dn_c + en_a$ where d and e are constants. If, now, $e/d = 0$, π_c depends only on the number of conspecific females,

while if $e/d = 1$ it depends only on the total number of females. Since π_c apparently decreases for $n_c = 40$, $n_a = 10$ and not for $n_c = 10$, $n_a = 40$, e/d seems to be a positive fraction less than one.

Stalker[3] has proposed an "isolation index" $\beta = (\pi_c - \pi_a)/(\pi_c + \pi_a)$. An equivalent but somewhat simpler index is $\gamma = \pi_a/\pi_c$. They may be interchanged by the formulas $\beta = (1 - \gamma)/(1 + \gamma)$ and $\gamma = (1 - \beta)/(1 + \beta)$. If only homogamic matings occur, $\beta = 1$ and $\gamma = 0$; if there is no discrimination, $\beta = 0$ and $\gamma = 1$; and if only heterogamic matings occur, $\beta = -1$ and $\gamma = \infty$. It is somewhat easier to interpret γ than β, as the probability of an alien female being inseminated is simply γ times the probability of a conspecific female being inseminated. We shall call γ the "isolation ratio."

We shall now consider the implications of the hypotheses, discussed above, about the manner in which the mating preferences become effective. If males do not mate with alien females unless conspecific females are unavailable, we should expect to find no inseminated aliens until nearly all the conspecific females have been inseminated. Since this is not so, we must look for some further explanation. It might be that only a small proportion of the females present are receptive at any one time. This seems rather farfetched; it might, however, be tested by seeing what happens when only one or two males are present. On the other hand, the hypothesis might be altered to give a hypothesis which we will call H1 by supposing that the probability π_a of an alien female being inseminated is not zero but has some low value when conspecific females are available and has a higher value when only aliens are available. If we also suppose that the presence of conspecific females excites males and makes them more likely to mate, we will have a situation where π_a has one low value in the absence of conspecific females, another low value when conspecific females are available, and a higher value when conspecific females have been available but are so no longer. The balance between these opposing tendencies would then determine whether π_a would be increased or decreased when the number of conspecific females n_c was increased from zero to, e.g., ten. Introducing a second subscript to refer to the experiment number, the question is whether π_{a3}/π_{a2} is greater than or less than one. The data give[4] $p_{a3}/p_{a2} = 1.23$, but this does not differ significantly from one. This comparison was repeated, using aged flies to decrease the heterogeneity as explained above. When these vials were tested for heterogeneity they gave $\chi^2 = 19.56$ with 28 degrees of freedom, for $P = 0.85$, so that these results are much more reliable. This time $p_{a9}/p_{a8} = 1.36$ and this is significant, with $\chi^2 = 4.81$ and $P = 0.03$.

Before discussing H1 further we will consider an alternative hypothesis, H2. We now suppose that the isolation ratio $\gamma = \pi_a(n_c, n_a, t)/\pi_c(n_c, n_a, t)$ is a constant independent of n_c, n_a, and t. Evidently this cannot hold for

all possible n_c, n_a and t; for example, if t is very large and n_c and n_a small, then π_c, π_a and hence γ will tend to 1; however γ may be essentially constant over a wide range of values. We shall not specify any definite forms for the functions π_c and π_a but merely suppose that in the xth experiment π_c has a fixed value π_{cx} and π_a has a fixed value π_{ax}. We temporarily ignore the heterogeneity of the data. Hypothesis H2 now reduces to H2′, where we assume that π_a/π_c has a common value γ_x for each vial of experiment x, and the hypothesis states that all the γ_x are equal. We represent the common value of the γ_x by γ.

Evidently γ_x involves a comparison between two kinds of females and hence cannot be computed for experiments 1 and 2, and therefore they cannot be used in testing H2′. For the other experiments, under the assumption only (i.e., when the γ_x need not be equal), the best estimates for π_{cx}, etc., are p_{cx}, etc., where p_{cx} is the proportion of conspecific females fertilized in all vials of the xth experiment; and the best estimate of γ_x is $g_x = p_{ax}/p_{cx}$. The g_x are given in table 1.

When H2′ is true, the p's are no longer the best estimates. We may now obtain π_{ax} from the relation that under H2′, $\pi_{ax} = \gamma\pi_{cx}$ but we still require joint estimates of $\gamma, \pi_{c3}, \pi_{c4}, \ldots, \pi_{c7}$. The method of maximum likelihood[5] leads to the set of six non-linear equations

$$\left.\begin{aligned} &\sum_{x=3}^{7}\left[\frac{F_{ax}}{\gamma} - \frac{\pi_{cx}(N_{ax}-F_{ax})}{1-\gamma\pi_{cx}}\right] = 0 \\ &\left[\frac{F_{cx}+F_{ax}}{\pi_{cx}} - \frac{N_{cx}-F_{cx}}{1-\pi_{cx}} - \frac{\gamma(N_{ax}-F_{ax})}{1-\gamma\pi_{cx}}\right] = 0 \quad (x = 3, 4, \ldots, 7). \end{aligned}\right\} \tag{1}$$

It is not practicable to solve these equations exactly, but a method of successive approximations gives the estimate $g = 0.1958$ and estimates of the π_{cx} which lead to the "expected values" that are given in table 1. The ordinary χ^2 test is not strictly applicable here because the constraints are non-linear. However the likelihood ratio statistic,[6] λ, is easily calculated and, for samples as large as this, $-2\log_e\lambda$ should be distributed approximately as χ^2 with 4 degrees of freedom.[7]

To compute $-2\log_e\lambda$ we need the expression

$$(-2\log_e 10)\sum_{x=3}^{7}[(F_{cx}+F_{ax})\log\pi_{cx} + F_{ax}\log\gamma_x + (N_{cx}-F_{cx})\log(1-\pi_{cx}) + (N_{ax}-F_{ax})\log(1-\gamma_x\pi_{cx})]. \tag{2}$$

(All logarithms are to the base 10 unless otherwise indicated.) Now let L_1 be the value of (2) when we replace γ_x by γ and then replace $\gamma, \pi_{c3}, \pi_{c4}, \ldots, \pi_{c7}$ by the numerical estimates obtained fom equations (1). Also, let L_2 be the value of (2) when we replace $\gamma_3, \ldots, \gamma_7, \pi_{c3}, \ldots, \pi_{c7}$, by the

sample values given in table 1. Then $-2 \log_e \lambda = L_1 - L_2$ We obtain the result $-2 \log_e \lambda = 9.0$, corresponding to a probability of about 0.061.

Discussion.—In nature, H2 would mean essentially that there is one probability of copulation when a male meets one of his own females and another (smaller) probability when he meets an alien female, while under H1 we have the added complication that his willingness to mate with the alien is reduced by the presence of receptive conspecific females. It should be noted that we have spoken throughout as if it were the male who did the choosing. This has been done solely for economy of words. Actually it is not known whether it is the males or the females or both which exercise the discrimination. It may be that the alien females reject the males, and that H1 means that the males are more likely to persevere if no conspecific females are present. The end result will be the same in any case. The important thing is that, under H2, π_a and π_c both decrease at the same rate, as n_c increases, while, under H1, π_a decreases more rapidly than π_c so that γ decreases. The data also require the assumption that males become excited if conspecific females are present, and then are more likely to mate with alien females. This additional effect could be superimposed on either H1 or H2, but would have to be more extreme to override the effect of H1. Since there is no other evidence in favor of H1, it seems that the simpler hypothesis H2 should be accepted. We have seen that, ignoring heterogeneity, the probability of the data arising by chance under H2′ is about 0.06. The heterogeneity would tend to increase the probability of more extreme values of the statistic $-2 \log_e \lambda$, and hence to further reduce the need of rejecting H2′. On the other hand, the heterogeneity would also tend to overshadow any real differences in the γ_r and hence make the test less sensitive. Table 1 shows on its face that the g_r do vary over a range of 2 to 1, but in an irregular manner rather than in the regular way which we would expect under H1. The actual variation in the γ_r could be even greater than this, but it is not likely that very large systematic changes occur. All things considered it seems best to accept the simpler hypothesis H2, since it is not contradicted by the data, and indeed seems to fit them better than H1. In any event the simple hypothesis that males only mate with alien females when their own females are unavailable seems untenable, and its modification, H1, is simply a more complicated version of H2.

Results with Two Strains of D. prosaltans.—Experiments were also carried out with the same number of flies as in experiments 1 to 7, but using flies of the Chilpancingo (Mexico) strain of *D. prosaltans* instead of *D. persimilis* and flies of the Bertioga (Brazil) strain of *D. prosaltans* instead of *D. pseudoobscura*. As shown by Dobzhansky and Streisinger,[8] males of both these strains exhibit a strong preference for mating with Chilpancingo females. Freshly hatched flies were kept together in an incubator at $25^1/_2$°C. for approximately 3 days, whereupon all the females were dissected and their

seminal receptacles examined for sperm. These experiments were performed in August–September, 1944. The results are summarized as experiments 1′ to 7′ in table 2.

TABLE 2

D. prosaltans. NUMBERS OF INSEMINATED FEMALES OBSERVED WHEN 10 CHILPANCINGO MALES WERE PLACED WITH n_c CHILPANCINGO FEMALES AND n_a BERTIOGA FEMALES

EXPT. NUMBER x	NO. OF CONSPECIFIC PER VIAL n_{cx}	NO. OF ALIEN PER VIAL n_{ax}	CHILPANCINGO FERTILIZED F_{cx}	CHILPANCINGO UNFERTILIZED $N_{cx}-F_{cx}$	CHILPANCINGO TOTAL N_{cx}	BERTIOGA FERTILIZED F_{ax}	BERTIOGA UNFERTILIZED $N_{ax}-F_{ax}$	BERTIOGA TOTAL N_{ax}	ISOLATION INDEX b_x	ISOLATION RATIO g_x
1′	10	0	57*	33	90					
			63.33*	36.67						
2′	0	10				14	86	100		
						14.00	86.00			
3′	10	10	72	27	99	2	95	97	0.945	0.0284
			72.73	27.27		2.06	97.94			
4′	20	10	116	58	174	1	86	87	0.966	0.0172
			66.67	33.33		1.15	98.85			
5′	40	10	156	240	396	1	97	98	0.950	0.0259
			39.39	60.61		1.02	98.98			
6′	10	20	57	22	79	3	152	155	0.948	0.0268
			72.15	27.85		1.94	98.06			
7′	10	40	64	32	96	12	381	393	0.912	0.0458
			66.67	33.33		3.05	96.95			

* For each experiment, the top line gives the observed number of flies and the bottom line gives the observed percentages.

In these experiments the observed frequencies of heterogamic fertilization are so low that their sampling fluctuations would overshadow any reasonable changes in the γ_x. All that can be said is that the data in experiments 3′ through 7′ do not contradict H2′. For such low γ values large numbers of vials would have to be used to obtain good results. The only significant result is the considerable increase of heterogamic matings when no conspecific females are present. We have $p_{a3}'/p_{a2}' = 0.147$, with $\chi^2 = 7.43$, using Yates' correction for continuity, and $P = 0.006$. This is probably a real effect, although it might be due to the heterogeneity between vials. If real, this effect is more consistent with H1 than with H2. It could be at least partly explained under H2 as follows. Suppose π_c and π_a depend on an "effective population density" with an added female affecting π_a in proportion to her chance of insemination; i.e., if, as suggested above, we consider π_a to depend on a linear combination, $n' = dn_c + en_a = d[n_c + (e/d)n_a]$, of n_c and n_a, then e/d will be equal to γ. Now if $\pi_a(n', t)$ decreases rapidly as n' increases from 0 to 2 or 4, and thereafter decreases less rapidly, then if γ is small, as it is here, 10 γ will be small, and increasing n_c from 0 to 10 will reduce π_a considerably. There is as yet no experimental evidence on the behavior of π_c and π_a for small n_c by which this conjecture could be tested.

Thus the difference in behavior in the two series of experiments need not necessarily imply different causal systems. However, it would not be surprising if they were different, since in one case we are dealing with sexual isolation between two species and in the other with a one-way sexual preference between two strains of a single, different species.

Summary.—Results obtained by placing *D. persimilis* males with varying proportions of *D. persimilis* and *D. pseudoobscura* females admit of the hypothesis that the ratio of the probability of a heterogamic mating to the probability of a homogamic mating is a fixed constant independent of these proportions. However, the possibility of some decrease in this ratio when many *D. persimilis* females are present cannot be rejected. Because of small numbers of heterogamic matings, similar experiments with two strains of *D. prosaltans* furnish little evidence on this point. For *D. persimilis-D. pseudoobscura*, heterogamic matings are significantly more frequent in vials containing 10 females of each of the two species than in vials containing 10 alien females only, while for *D. prosaltans* they are significantly less frequent.

* Experimental data by Th. Dobzhansky, mathematical analysis by H. Levene.

[1] Mayr, E., and Dobzhansky, Th., these PROCEEDINGS, **31**, 75–82 (1945).

[2] For simplicity, we shall use the terms "conspecific" and "alien" regardless of whether the flies differ as to species (*D. persimilis* and *D. pseudoobscura*) or merely as to strain (geographic strains of *D. prosaltans*).

[3] Stalker, H. D., *Genetics*, **27**, 238–257 (1942).

[4] Throughout this paper Greek letters are used for population values (i.e., values characteristic of a given experimental set-up) and the corresponding Roman letters are used for estimates based on the sample of flies observed. Thus the best estimate of π_{ci} is $p_{ci} = F_{ci}/N_{ci}$.

[5] Fisher, R. A., *Statistical Methods for Research Workers*, London, Sec. 53 ff., 2nd (1928) and later editions.

[6] Neyman, J., and Pearson, E. S., *Biometrika*, **20A**, 175–240 and 263–294 (1928).

[7] Wald, A., *Trans. Am. Math. Soc.*, **54**, 428–482 (1943).

[8] Dobzhansky, Th., and Streisinger, G., these PROCEEDINGS, **30**, 340–345 (1944).

THE THREE-DIMENSIONAL SHAPES OF BUBBLES IN FOAMS

BY EDWIN B. MATZKE

DEPARTMENT OF BOTANY, COLUMBIA UNIVERSITY

Communicated July 23, 1945

The three-dimensional shapes of bubbles in foams have been discussed by scientists in diverse fields for nearly three centuries, but apparently exact studies have not heretofore been made.

Even in his *Micrographia*, published in 1665, Robert Hooke,[1] the first

describer of cells, alluded to "frothy bodies" and "congeries of very small bubbles," while his contemporary, Nehemiah Grew,[2] in 1682 drew more direct comparisons between plant cells and the "Froth of Beer or Eggs." Ever since, the similarity between cells in tissues and bubbles in foam has been stressed. Thus, Holman, as reported by Child,[3] made a foam which he compared with cells in masses, while Robert[4] published a striking series of photographs of groups of bubbles simulating the early cleavage stages of the gasteropod *Trochus*. Berthold[5] and Errera[6] both showed how cells and soap films have certain features in common. However, until Lewis[7, 8, 9] began his painstaking researches on the three-dimensional shapes of cells in tissues, no precise information was available on the shapes of either cells or bubbles in foam, even though they had often been compared.

If biologists have been more interested than other scientists in this topic, they are by no means the only ones who have discussed it. Lord Kelvin,[10, 11] the distinguished mathematician and physicist, who was strongly influenced by the scholarly volumes on soap films written by the blind Belgian physicist Plateau,[12] published several essays on "the division of space with minimum partitional area," "homogeneously," "into equal and similar parts," "all sameways oriented" and under conditions of "stable equilibrium." Kelvin solved the problem of division of space into equal and similar units with stable angles and each with maximum volume and minimum surface by describing a 14-faced figure, the minimal tetrakaidecahedron, with 8 undulating hexagonal faces and 6 quadrilateral faces with the sides curved. This figure has stable angles. A similar plane-faced tetrakaidecahedron, with 8 faces which are regular hexagons and 6 which are squares, does not have completely stable angles. Kelvin's figures have been widely accepted, and under ideal conditions with *perfect spacing*, they may well be realized.

Sir James Dewar[13, 14, 15] made extensive studies of soap films and bubbles. Separate volumes have been written on the subject by Lawrence[16] and by Boys,[17] while Courant and Robbins[18] stress the importance of soap-film studies in the solution of certain mathematical problems.

But the only one who has seriously tried to examine the three-dimensional shape of bubbles in an actual foam is Desch,[19] who became interested in the problem through its bearing on the form of crystal grains in solidifying metals, as in marine propellers and in crucible steel. Desch's data are not extensive, and are somewhat at variance with the results presented below.

This same topic has been of interest also to students of the alveolar structure of protoplasm. Bütschli[20] concluded that protoplasmic alveoli were dodecahedral, while Seifriz[21] thinks they may be 12-faceted or 14-faceted.

It is obvious, therefore, that the three-dimensional shape of bubbles in foam has been of interest and of philosophical import to physicists, mathe-

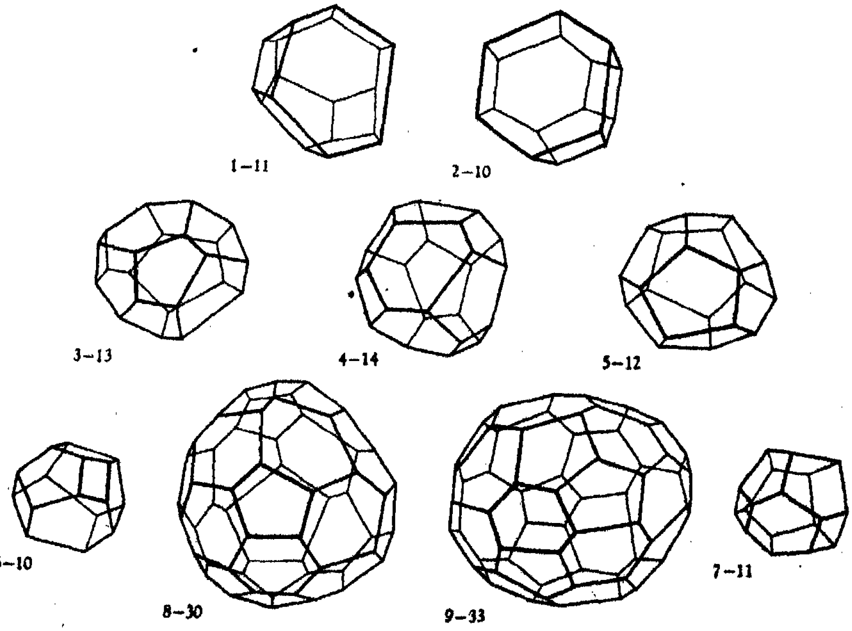
maticians, metallurgists, cellular biologists and students of protoplasmic structure. It is also evident that no one heretofore has succeeded in looking into a large mass of soap bubbles and tabulating the number and kinds of faces on each.

In order to accomplish this, at least under one set of conditions, a series of studies on the three-dimensional shapes of bubbles in foams have been carried out, using a technique different from that of previous investigators. The soap solution found most satisfactory was essentially that of Johnston,[22] and consisted of the following parts: triethanolamine oleate 7.5 g., glycerin 34 g., and distilled water 58.5 g. These substances were shaken together thoroughly at intervals for several days, and then allowed to stand for at least several weeks.

Each of the bubbles used was made separately by means of a Yale tuberculin syringe graduated to $^1/_{100}$ cc. If the bubble was to be $^1/_{10}$ cc. in volume, the plunger was set at $^1/_{10}$ cc., the tip of the syringe was dipped into the soap solution, and the bubble was blown by pressing down the plunger. Each bubble was placed separately into a cylindrical dish, having an inside diameter of 6 cm. and an inside height of 6.5 cm. The inside of the dish was moistened beforehand with the soap solution. A glass plate was used as a cover. After each bubble was "blown," excess soap solution was removed by wiping the tip of the syringe on cheesecloth. For most of the work the dish was not completely filled, so that the bubbles did not come in contact with the cover, even when temperature changes resulted in expansion. The dish was rotated from time to time, to keep the solution distributed as uniformly as possible and to allow the bubbles to slip and become readjusted. If the bubbles were to be kept for some time, the dish was rotated on its long axis on a clinostat. Although the bubbles would last for days, they were always studied as soon as possible after they were made, usually on the same day. Even with rotation, some breaking did occur, so that the mass of foam cells was not completely static, though a state of relatively stable equilibrium was achieved. When bubbles of varying sizes were used, the large ones tended to increase in volume, and the small ones to decrease, over a period of days. If breaking occurred, all the bubbles in the dish were discarded. Although only 1300 bubbles were used in the tabulations below, a total of approximately 57,200 bubbles were made, each individually, and placed separately into the dish in the course of the following experiments.

The bubbles, in the dish, were examined under a Spencer binocular dissecting microscope with 1X paired objectives and 6X paired oculars. If the films were thin, it was possible to focus through a dish of these bubbles and with the dissecting microscope to single out any individual one in a mass of approximately 1800, and to record exactly the number and kinds of faces.

The illustrations shown in figures 1 to 9 were made with a Leitz camera lucida attached to the dissecting microscope. The drawings were done by plotting the corners of each face and connecting them with straight lines. In this way the beautiful, delicate curvatures of the faces and edges, by which stability of angles is achieved, are not shown; this method was adopted for expediency, since to show these curvatures would be exceedingly difficult; the numbers and kinds of faces are truly represented.



FIGURES 1–9

Camera Lucida drawings of bubbles. Xc 3. The second number under each figure represents the total number of faces. Curvatures of faces and edges are not shown in the drawings. Quadrilateral faces are designated as Q, pentagonal as P, hexagonal as Hx, heptagonal as Hp. The combinations of faces in each are as follows: Fig. 1.—11 faces, 3 Q, 6 P, 2 Hx. Fig. 2.—10 faces, 4 Q, 4 P, 2 Hx. Fig. 3.—13 faces, 1 Q, 10 P, 2 Hx. Fig. 4.—14 faces, 2 Q, 8 P, 4 Hx. Fig. 5.—12 faces, 12 P. Fig. 6.—10 faces, 3 Q, 6 P, 1 Hx. Fig. 7.—11 faces, 2 Q, 8 P, 1 Hx. Fig. 8.—30 faces, 13 P, 16 Hx, 1 Hp. Fig. 9.—33 faces, 14 P, 17 Hx, 2 Hp.

In the first set of experiments, bubbles of uniform volume were used. The dish was filled, nearly to the top, with bubbles of $^1/_{10}$ cc. or $^1/_{20}$ cc. With each filling, bubbles of one volume only were utilized—they were not mixed. Since almost identical results were obtained with volumes of $^1/_{10}$ cc. and $^1/_{20}$ cc., all the data are combined in table 1.

It is immediately evident from table 1 that 400 peripheral bubbles—including 200 in contact with the walls of the cylinder and 200 on the free upper surface—had an average of 10.99 faces, and that the range was from 7 to 14. Bubbles with 11 facets occurred more commonly than any other kind. In table 1 it can be seen also that there were more pentagonal faces (2338) in these peripheral bubbles than all other kinds combined (2057).

TABLE 1

NUMBER AND KINDS OF FACES OF PERIPHERAL AND CENTRAL BUBBLES OF UNIFORM VOLUME

NO. OF FACES	NO. OF BUBBLES PERIPHERAL	NO. OF BUBBLES CENTRAL
7	2	...
8	13	...
9	20	...
10	88	...
11	154	2
12	83	73
13	33	179
14	7	218
15	...	106
16	...	20
17	...	2
	400	600
Average	10.99	13.70

KINDS OF FACES	PERIPHERAL	CENTRAL
Triangular	11	..
Quadrilateral	1252	866
Pentagonal	2338	5503
Hexagonal	720	1817
Heptagonal	73	35
Octagonal	1	..
Totals	4395	8221

The 400 peripheral bubbles here recorded showed 43 different combinations of faces. The combination occurring with greatest frequency had 3 quadrilateral, 6 pentagonal and 2 hexagonal faces (3 Q, 6 P, 2 Hx). It was found 67 times among the 400 peripheral bubbles tabulated, and is shown in figure 1. Normally the outer face was hexagonal, and there were 6 lateral and 4 basal contacts. The second most common combination in the peripheral bubbles (4 Q, 4 P, 2 Hx) is shown in figure 2; it occurred 47 times. Although certain of these combinations occurred much more frequently than others—9 of the 43 were found only once—still no one of these may be considered as "the type."

In addition to those on the surface, 600 "central" bubbles were examined, and the data are recorded in table 1. By "central" are meant those that were separated from the top, bottom or sides of the cylinder by at least three others. An inspection of the table reveals that the number of faces of the central bubbles was higher than that of the peripheral, that those with 14 contacts were most numerous, that the range was from 11 to 17,

and that the average for 600 was 13.70. It is further obvious that a great majority of the faces (5503 of 8221) were pentagonal.

The 600 central bubbles here tabulated were found in 36 different combinations of faces. The combination found most frequently was a 13-faceted foam cell with 1 quadrilateral, 10 pentagonal and 2 hexagonal contacts. It occurred 118 times among the 600, and is illustrated in figure 3. The second and third most common combinations—1 Q, 10 P, 3 Hx, and 2 Q, 8 P, 4 Hx (Fig. 4)—were both 14-faceted foam cells, and they were found 73 and 64 times, respectively. The fourth most common combination of the 36 was the pentagonal dodecahedron (Fig. 5) with 12 faces all pentagons; there were 50 of these. But, as in the peripheral bubbles, no one of these combinations may be singled out as "the type," and although the average number of contacts is close to 14, Kelvin's tetrakaidecahedra, with 8 hexagonal and 6 quadrilateral faces, are not realized under the conditions present here.

In a second set of experiments, bubbles of two different volumes were mixed together in known proportions. Small bubbles were made with a volume of 0.05 cc., and large bubbles with a volume of 0.4 cc. The results are summarized in table 2. In experiment 1, table 2, 2 small bubbles were made for each large one, so that when the dish was nearly filled and ready for examination, it contained twice as many small bubbles as large ones. In experiment 2, table 2, 8 small bubbles were put into the dish for each large one; in experiment 3, table 2, the ratio was 32 small to 1 large. Fifty small bubbles and 50 large ones were then studied in each of these 3 experiments, and the data recorded. Only "central" small and large bubbles were used—that is, none was tabulated unless it was at least the fourth one from the top, bottom or sides of the dish. During the course of these three experiments, in which 150 small and 150 large bubbles were studied, the dish was filled 19 times.

An analysis of table 2 reveals that when small and large bubbles are associated together, the number of contacts of the small ones is less than that of bubbles of uniform volume (13.70), while the number of contacts of the large ones is increased. Thus in experiment 1, table 2, the average of the small bubbles was 9.68, while that of the large ones was 20.42. Furthermore, increasing the ratio of small to large bubbles brought the number of contacts on the small ones closer to 13.70, and as the ratio of small to large increased, the number of contacts on the large bubbles became higher, within the limits of these experiments. On the basis of the data in table 2 and of the ratios of small to large bubbles used in the three experiments (that is, 2:1, 8:1, and 32:1), the average number of contacts of all the central bubbles (large and small) in the dish at one time can be computed; it is found to be 13.26 for experiment 1, 13.62 for experiment 2 and 13.71 for experiment 3. Therefore, when bubbles of two different volumes are asso-

ciated in varying but known proportions, the number of contacts is altered in definite ways.

TABLE 2

Number and Kinds of Faces of Central Bubbles of Two Different Volumes Associated in Varying Proportions

No. of faces	Number of bubbles — Small bubbles Expt. 1	Expt. 2	Expt. 3	Large bubbles Expt. 1	Expt. 2	Expt. 3	Totals
7	1	...	...	...	...	...	1
8	4	...	...	...	...	...	4
9	17	1	...	...	...	...	18
10	18	3	...	...	...	...	21
11	8	14	1	...	...	...	23
12	2	18	16	...	...	...	36
13	..	10	16	...	...	...	26
14	..	4	9	...	...	...	13
15	..	...	6	...	...	...	6
16	..	...	1	...	...	...	1
17	..	...	1	1	...	...	2
18	..	...	...	5	...	...	5
19	..	...	...	6	...	...	6
20	..	...	...	14	...	...	14
21	..	...	...	12	...	...	12
22	..	...	...	10	...	...	10
23	..	...	...	0	...	...	0
24	..	...	...	2	4	...	6
25	..	...	...	...	7	...	7
26	..	...	...	...	6	...	6
27	..	...	...	...	10	2	12
28	..	...	...	...	8	5	13
29	..	...	...	...	5	10	15
30	..	...	...	...	9	13	22
31	..	...	...	...	1	12	13
32	..	...	...	...	...	7	7
33	..	...	...	...	...	1	1
	50	50	50	50	50	50	300
Average	9.68	11.90	13.20	20.42	27.34	30.06	

Kinds of faces	Small bubbles Expt. 1	Expt. 2	Expt. 3	Large bubbles Expt. 1	Expt. 2	Expt. 3	Totals
Quadrilateral	159	118	96	115	57	36	581
Pentagonal	282	370	412	498	625	624	2811
Hexagonal	43	101	148	288	557	749	1886
Heptagonal	...	6	4	113	118	92	333
Octagonal	...	...	...	6	9	2	17
Nonagonal	...	...	...	1	1	..	2
	484	595	660	1021	1367	1503	5630

From table 2 it can be seen that pentagonal faces occurred more commonly than any other kind, constituting very nearly half the total number. Hexagons were also fairly abundant.

Varying the volumes and ratios of the bubbles results in an increase in the number of combinations of faces. Thus the 150 small bubbles were found in 35 different combinations, while the 150 large ones occurred in 105 combinations. Among the small bubbles the combination found most often (16 times) had 10 faces, 3 Q, 6 P, 1 Hx; it is illustrated in figure 6. Of the two second most abundant, both of which occurred 12 times, one had 11 faces, 2 Q, 8 P, 1 Hx, figure 7, and the other had 13 faces, 1 Q, 10 P, 2 Hx. This latter is also the commonest combination in central bubbles of uniform volume (Fig. 3). The large foam cell occurring most frequently (7 times) is shown in figure 8; it had 30 faces, 13 P, 16 Hx, 1 Hp. One bubble with 33 faces was found—the largest number recorded—14 P, 17 Hx, 2 Hp (Fig. 9). Here again, no one "type" is present in either the large or the small bubbles. Pentagonal faces occur frequently, hexagonal somewhat less so, but in no uniform pattern.

A consideration of the data presented in tables 1 and 2 leads to certain definite conclusions:

1. The peripheral foam cells of a mass of bubbles of uniform volume have an average of 11 faces. Lewis[23] had previously recorded 11-faceted epidermal cells.

2. The central bubbles in a foam, in which all are of approximately equal volume, have an average of nearly 14 faces (13.70). But pentagonal faces predominate, so that Kelvin's tetrakaidecahedra with 8 hexagonal and 6 quadrilateral faces are not realized under these conditions. This does not mean that Kelvin was incorrect, for he was interested in the theoretical division of space homogeneously into equal and similar units, similarly oriented and having stable angles. When a large number of bubbles of equal volume are placed into a cylindrical dish, so that they are free to glide and adjust themselves, they become arranged in a mass of somewhat irregular foam cells, approximating, at least, the equilibrium conditions outlined by Plateau, but Kelvin's "ideal" figure is not achieved. This is at variance with the conclusion presented in Sir D'Arcy Thompson's[24] scholarly volume.

3. Comparisons may be made between the data presented above and similar data on lead shot of uniform volume (Marvin[25]) and of varying volumes (Matzke[26]) compressed to eliminate interstices. In such a system, starting with spherical shot, the effects of surface energy in subsequent changes in shape are obviously of minor importance, in contrast to the conditions prevailing in soap films. The range in number of contacts, in kinds of faces, and especially in number of combinations of faces is more restricted in the foam cells than in the compressed lead shot.

4. Detailed comparisons will be published elsewhere between the complete data on soap films and cells in plant and animal tissues, described by Lewis, Marvin,[27, 28] Higinbotham,[29] Dodd[30] and Hulbary.[31] It may be

noted here, however, that cells in essentially undifferentiated tissues are intermediate in many respects between bubbles in foam on the one hand and compressed lead shot on the other. And frequently the cells approach the bubbles in foam distinctly more closely than they do the compressed lead shot. This is interpreted as indicating that surface forces are operative when the delicate cell walls are laid down in or near the plant meristems, that surface forces are of relatively less moment in cellular organization than in soap films, and that they constitute one of the factors, but only one, in three-dimensional cell shape determination.

[1] Hooke, R., *Micrographia*, London, 1665.

[2] Grew, N., *The Anatomy of Plants*, 2nd ed., London, 1682.

[3] Child, J. M., *Jour. Franklin Inst.*, **80**, 252–253 (1880).

[4] Robert, A., *Arch. de Zoologie Exp. et Gén.*, 3 ser., **10**, 269–538 (1902).

[5] Berthold, G., *Studien über Protoplasmamechanik*, Leipzig, 1886.

[6] Errera, L., *Bot. Centralbl.*, **34**, 395–398 (1888).

[7] Lewis, F. T., *Proc. Amer. Acad. Arts and Sci.*, **58**, 537–552 (1923).

[8] Lewis, F. T., *Ibid.*, **61**, 1–34 (1925).

[9] Lewis, F. T., *Ibid.*, **68**, 251–284 (1933).

[10] Kelvin, Lord, *Phil. Mag.*, 5s, **24**, 503–514 (1887).

[11] Kelvin, Lord, *Proc. Roy. Soc. London*, **55**, 1–16 (1894).

[12] Plateau, J., *Statique expérimentale et théorique des liquides soumis aux seules forces moléculaires*, Paris, 1873.

[13] Dewar, J., *Collected Papers of Sir James Dewar*, Vol. 2, 1154–1176 (1916).

[14] Dewar, J., *Collected Papers of Sir James Dewar*, Vol. 2, 1176–1206 (1917).

[15] Dewar, J., *Collected Papers of Sir James Dewar*, Vol. 2, 1206–1241 (1918).

[16] Lawrence, A. S. C., *Soap Films*, London, 1929.

[17] Boys, C. V., *Soap-bubbles and the Forces Which Mould Them*, London, 1890.

[18] Courant, R., and Robbins, H., *What Is Mathematics?* New York, 1941.

[19] Desch, C. H., *Jour. Inst. Metals*, **22**, 241–263 (1919).

[20] Bütschli, O., *Investigations on Microscopic Foams and on Protoplasm.* Trans. by E. A. Minchin, London, 1894.

[21] Seifriz, W., *Protoplasm*, New York, 1936.

[22] Johnston, B., *Civil Eng.*, **5**, 698–701 (1935).

[23] Lewis, F. T., *Science*, **86**, 609–611 (1937).

[24] Thompson, D'Arcy, *On Growth and Form*, New ed., 1942, p. 552.

[25] Marvin, J. W., *Amer. Jour. Bot.*, **26**, 280–288 (1939).

[26] Matzke, E. B., *Ibid.*, **26**, 288–295 (1939).

[27] Marvin, J. W., *Ibid.*, **26**, 487–504 (1939).

[28] Marvin, J. W., *Ibid.*, **31**, 208–218 (1944).

[29] Higinbotham, N., *Ibid.*, **29**, 851–858 (1942).

[30] Dodd, J. D., *Ibid.*, **31**, 120–127 (1944).

[31] Hulbary, R. L., *Ibid.*, **31**, 561–580 (1944).

GENETIC ANALYSIS OF THE INDUCTION OF TUMORS BY METHYLCHOLANTHRENE. XI. GERMINAL MUTATIONS AND OTHER SUDDEN BIOLOGICAL CHANGES FOLLOWING THE SUBCUTANEOUS INJECTION OF METHYLCHOLANTHRENE

By Leonell C. Strong*

From the Department of Anatomy, Yale University School of Medicine

Communicated July 19, 1945

The injection into mice, or the application onto the skin surface, of methylcholanthrene or any one of a large number of carcinogens is followed by a great variety of neoplasms depending upon variations of experimental technique. The variation in the types of cancers obtained with the injection of methylcholanthrene has been considerably increased by the use of hybrid mice (from the F_2–F_{18} generations following an outcross) when the normally expected local tumors at the site of injection have been partially suppressed or completely inhibited through genetic selection toward resistance to such locally appearing tumors.[1]

The actual manner by which a carcinogen initiates normal tissues or cells to become cancerous is not clear. It is generally accepted, however, that the carcinogen or one of its metabolites acts directly upon cells in bringing about the neoplastic change. The observation of Mottram that carcinogens cause chromosomes to lag in the equatorial plate during mitosis is highly suggestive that nuclear changes may be involved in the origin of cancer.[2] The recent observation of the author that the untreated descendants of mice developing cancer of the stomach following the subcutaneous injection of methylcholanthrene continue to develop the same type of gastric cancer spontaneously is evidence that this particular neoplastic condition may have been hereditarily established through at least four generations by the effect of methylcholanthrene upon the germ plasm.[3] Linkage tests have demonstrated that susceptibility to gastric carconima is determined by the effect of a gene on the "brown tagged" chromosome.[3]

If it could be demonstrated that germinal changes (mutations) can be induced by methylcholanthrene, it would be strong presumptive evidence that perhaps the action of methylcholanthrene (and the other carcinogens) in bringing about cancer is the induction of a somatic mutation. This explanation is more consistent than the assumption that methylcholanthrene has one effect on the germ plasm (induction of mutations) and still another totally unrelated effect upon somatic tissue in converting them into cancer. Some evidence that mutations affecting susceptibility to tumors induced by methylcholanthrene do occur in mice following the injection of the carcinogen has already been published.[4]

It is the purpose of this paper to present evidence that germinal mutations

and other suddenly appearing biological characteristics arise in the descendants of mice that have received a subcutaneous injection of methylcholanthrene and that these sudden changes arise more frequently than are to be expected by chance alone.

Results.—One hundred fifty-five F_1 mice have been produced by a cross between C_{57} (homozygous for dark eye, black, non-agouti and self) and one subline of the NHO descent (proved to be homozygous for dark eye, brown, non-agouti and self). The above outcross was made in both directions. In each case only the NHO parent was injected with methylcholanthrene. Of these, 154 had the expected dark eye, black, non-agouti and self characters, while one was dark eye, black, non-agouti and spotted (proved to be dominant in inheritance). The new variant of spotting involved only the ventral surface of the body. Thus a dominant mutation of spotting occurred once in 155 mice. The author has not had dominant spotting (black eye white) in his laboratory for 25 years. It is also possible, although not definitely proved, that this new dominant spotting is not the same as the old black-eyed white character.

In all, thirteen color mutations have been obtained in the direct descendants of mice receiving methylcholanthrene. These are: (1) brown to black (dominant change), (2, 3, 4, 5) four pink eye mutants from dark eye (recessive inheritance), (6) color to albinism (recessive), (7) spotting involving only the ventral surface of the mouse discussed above (dominant), (8) a second type of spotting involving the nose and ventral surface only (dominant), (9) a third type of spotting involving only the belt region of the mouse (inheritance uncertain), (10) silvering (precocious graying evident upon the appearance of the first hair at 5–7 days of age) (inheritance uncertain), (11) permanent wave (inheritance uncertain), (12) piebald to self (dominant) and (13) gray belly to non-agouti in a form resembling black and tan. The thirteen new variants listed here are all proved to be *bona fide* mutations, based upon the observation of the subsequent appearance of the new forms in the lineal descendants of the mouse which first showed the character. The type of inheritance involved, determined by an analysis of the data obtained by an outcross to mice of an unrelated strain, has not been made in all cases—hence the use of the expression "type of inheritance uncertain."

Seven of these color mutations are repetitions of characters present in the author's laboratory, while six are new ones never having been present previous to the injection of methylcholanthrene.

The mutation rate is difficult to determine. For example, approximately 12,000 mice have been injected with 1 mg. of methylcholanthrene. Both parents of the NHO descent have been injected with the carcinogen for fourteen generations. The restrictions of space and time do not permit the keeping of all descendants of the experimental animals. In dealing with

hybrid mice as was done in this case, it is necessary to exclude all recessive variants that may have been due to segregation. A careful analysis of the data indicates that in these experimental animals in one restricted group the mutation rate was approximately 1 in 400[3]. In the controls over a period of 27 years, the mutation rate involving coat color changes has been approximately 1 in 26,000 mice.[3]

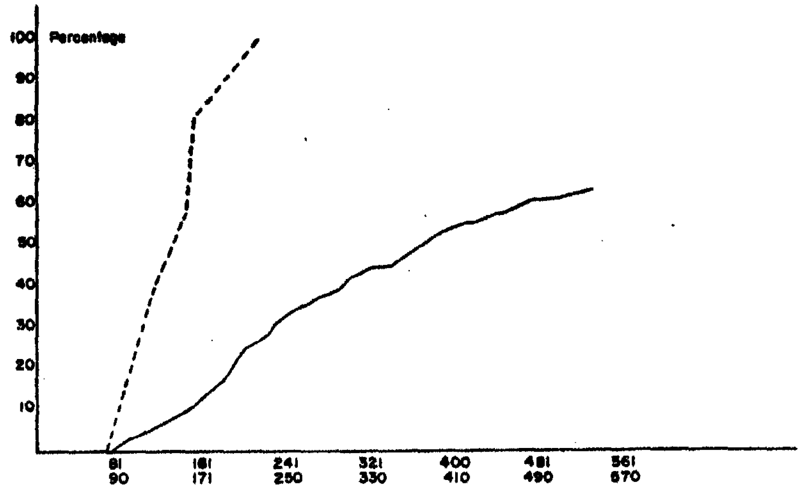


CHART ONE

Data on the altered susceptibility to local tumors induced by methylcholanthrene are given in this chart. Data on the original brown non-agouti subline are given on the solid line; data for the mutated black non-agouti subline are given on the dash line. Time in days is expressed along the base line; percentage incidence of induced local tumors is given on the vertical; that is, the data, with advancing time, are cumulative.

Two sudden changes, other than color changes and alterations affecting cancer susceptibility[4] have also arisen in the descendants of methylcholanthrene-treated animals. These are (1) the establishment of precocious sexual activity—the production of first litters by 42 days of age in one subline as compared to 78–85 days, the average in the original ancestral stocks which gave rise to them and (2) the establishment of a subline which has mice that give rise to extremely large first litters (average 10.2 young) as compared to 5.6 young in first litters of mice of the ancestral stocks.

Two other biological changes have also occurred. Whether they arose suddenly is not evident. These are (1) a more rapid growth rate in early life and (2) the attainment of adult body weight of 55–60 g. as compared to 26–28 g. for mice of the ancestral stocks.

The derived color mutant sublines either (1) retain the degree of cancer susceptibility (latent period measured from the time of injection to the origin of the local tumors) possessed by the ancestral stock mice that gave rise to them or (2) suddenly acquire a new susceptibility to induced tumors far in excess of the original suscpetibility (Chart 1). The impression is gained that many new biological characteristics are being acquired at approximately the same time. Thus the conclusion is reached that the effect of methylcholanthrene upon the germ plasm at any one time is either (1) the induction of a single gene mutation or (2) changes far in excess of what are to be expected by a single point mutation.

Summary.—Thirteen mutations involving coat color have appeared in the untreated descendants of mice receiving methylcholanthrene. The first mutant obtained occurred in a subline of the NHO strain which had been subjected to methylcholanthrene in both parents for twelve generations. An increased growth rate, the attainment of a larger body size and weight, precocious sexual activity, large first litters and increased susceptibility to induced local tumors by methylcholanthrene have also been obtained. These suddenly appearing color mutations have occurred either alone or in association with one or more of the changed biological characteristics enumerated above. Thus there is evidence that methylcholanthrene has affected the germ plasm by bringing about germinal or point mutations and perhaps other undetermined effects. It is highly probable, therefore, that methylcholanthrene may also bring about malignancies in tissues by causing somatic mutations to arise in them.

* This experiment has been made possible by grants from the Jane Coffin Childs Memorial Fund for Medical Research and the Anna Fuller Fund.

[1] Strong, L. C., "Genetic Nature of the Constitutional States of Cancer Susceptibility and Resistance in Mice and Men," *Yale Jour. Biol. and Med.*, **17** 289–299 (1944).

[2] Mottram, J. C., "Some Effects of Cancer-producing Agents on Chromosomes," *Brit. Jour. Exper. Path.*, **15** 71–73 (1934).

[3] Strong, L. C., "Genetic Analysis of the Induction of Tumors by Methylcholanthrene. IX. Induced and Spontaneous Adenocarcinomas of the Stomach in Mice," *Jour. Nat. Cancer Inst.*, **5** 339–362 (1945).

[4] Strong, L. C., "Genetic Analysis of the Induction of Tumors by Methylcholanthrene VIII. Two Mutations Arising in Mice Following Injection of Methylcholanthrene," *Arch. Path.*, **39** 232–236 (1945).

DAMPING OF EPIDEMIC WAVES

By Edwin B. Wilson and Jane Worcester

Harvard School of Public Health

Communicated July 12, 1945

Soper[1] stated: "In the foregoing theory . . . infecting power was supposed to be instant, at the termination of a certain incubation period, and it appears that with such a law in operation, an initial upset of steady conditions of prevalence will be followed by epidemic waves that propagate themselves indefinitely, without diminution of amplitude." And he went on to develop another theory in which infecting power begins at the instant a person becomes infected and is spread over time but with a diminishing incidence; on this assumption he finds that there is damping of the epidemic waves. His arguments, be it noted, assume that the disturbance from the steady state may be treated as infinitesimal.

The first conclusion, namely, that, under the assumption that there is a fixed incubation period, there will be no damping, is incorrect, and its incorrectness derives from a particular one of the approximations Soper makes. He starts out with the equations

$$C(t) = (S/m)\, C(t - \tau), \qquad C = A - dS/dt \tag{1}$$

where C is the case rate, m a constant, S the number of susceptibles and A the rate of accession of susceptibles supposed steady, and he remarks that since the change in S is usually small in the unit interval (the incubation period τ) we may write these equations as

$$C(t + \tau/2) = (S/m)\, C(t - \tau/2), \qquad C = A - dS/dt. \tag{2}$$

This change in effect removes the lag due to the incubation period and therefrom is derived the absence of damping in the wave.

If we use D as a symbol of differentiation and $e^{hD}f$ as the symbol for Taylor's expansion and if we set $S = m + y$, with y small so that higher powers are to be neglected (1) and (2) become, respectively,

$$[\tau D(1 - e^{-\tau D}) + A\tau/m]y = 0, \tag{1'}$$

$$\left[\frac{\tau D}{2} \sinh \frac{\tau D}{2} + \frac{A\tau}{4m}\right]y = 0. \tag{2'}$$

The second equation has a negative root for D^2 and hence an undamped period in the neighborhood of $P = 2\pi\sqrt{m\tau/A}$. The first equation does not have a pure imaginary solution for D, but a solution with a negative real part leading to damped harmonic waves of the approximate form

$$y = e^{-At/4m} \cos \sqrt{A/m\tau}\; t$$

with period $P = 2\pi\sqrt{m\tau/A}$ (approximately[2]) and damping per period of $e^{-\pi/2\cdot\sqrt{A\tau/m}}$.

We may give the solution for a more general equation in which the generalized form of the law of mass-action is used and not only an incubation period τ but a duration σ of infectiveness (both constant) are assumed. The equation in S is then[3]

$$A - \left(\frac{dS}{dt}\right) = \left(\frac{S}{m}\right)^p [A + S(t - \tau - \sigma) - S(t - \tau)]$$

and if $S = m + y$ with y infinitesimal we have

$$\left[D + \frac{1}{\sigma} e^{-(\tau+\sigma)D} - \frac{1}{\sigma} e^{-\tau D} + \frac{Ap}{m}\right]y = 0.$$

If we let $a = \tau + \sigma/2$ we have

$$\left[aD^2 - \left(\frac{a^2}{2} + \frac{\sigma^2}{24}\right) D^3 + \left(\frac{a^3}{6} + \frac{a\sigma^2}{24}\right)D^4 - \ldots + \frac{Ap}{m}\right]y = 0.$$

When Ap/am is small the solution is approximately[4]

$$D = \frac{1}{4}\frac{pA}{m}\left(1 + \frac{\sigma^2}{12a^2}\right) + i\sqrt{\frac{pA}{ma}} \tag{3}$$

and from this the period and damping can be read off.

For the case of Soper's distributed infectiousness we should write in our notation (with $p = 1$ in the law of mass-action)

$$A - \frac{dS}{dt} = \frac{S}{m}\int_{-\infty}^{t} \left(A - \frac{dS}{dt}\right)_{\xi} e^{(\xi - t)/\tau}\, d\xi$$

so that new cases at time $\xi < t$ are effective at time t only in proportion[5] $e^{(\xi - t)/\tau}$. The mean value of the duration of infectiveness is then

$$\int_{-\infty}^{0} \xi e^{\xi/\tau} \div \int_{-\infty}^{0} e^{\xi/\tau}\, d\xi = \tau.$$

Now it so happens that[6]

$$\int_{-\infty}^{t} e^{(\xi - t)/\tau} f(\xi) d\xi = \frac{1}{1 + \tau D} f.$$

Hence the equation with the generalized mass law would be

$$A - \frac{dS}{dt} = \left(\frac{S}{m}\right)^p \left[A - \frac{D}{1 + D} S\right]$$

and for $S = m + y$, with y infinitesimal,

$$\left[D - \frac{D}{1 + \tau D} + \frac{Ap}{m}\right] y = 0 \quad \text{or} \left[D^2 + \frac{Ap}{m} D + \frac{Ap}{m\tau}\right] y = 0.$$

In this case the value of D may be obtained in finite form as

$$D = -\frac{Ap}{2m} + i\sqrt{\frac{Ap}{m\tau}}\sqrt{1 - \frac{Ap\tau}{4m}}.$$

If compared with (3) taking $\sigma = 0$ and $a = \tau$ we see that the waves have approximately the same period whether we use a fixed incubation period τ and an infinitesimal duration of infectiousness or whether we use an infinitesimal incubation period with a period of infectiousness which averages to be τ, it being understood that the infectiveness dies out exponentially. The former case has, however, half the damping of the latter instead of none at all as Soper states.

It is of further interest in connection with stepwise calculations to note that one may not replace derivatives in differential equations with differences without becoming liable to modify the indicated damping. For example, in illustrating his theory numerically Soper replaces

$$\frac{dx}{dt} = a - a\frac{x}{as}\frac{y}{a\tau}, \qquad \frac{dy}{dt} = a\frac{x}{as}\frac{y}{a\tau} - a\frac{y}{a\tau}$$

for purposes of calculation by

$$\Delta x = a - a\frac{x}{as}\frac{y}{a\tau}, \qquad \Delta y = a\frac{x}{as}\frac{y}{a\tau} - a\frac{y}{a\tau}.$$

If we set $x = as(1 + u)$, $y = a\tau(1 + v)$, with u and v infinitesimal the differential equations, of the infinitesimal epidemic (as found by Soper) are

$$s\frac{du}{dt} = -u - v, \qquad \tau\frac{dv}{dt} = u,$$

the solution of which leads to the equation for D

$$\begin{vmatrix} sD + 1 & 1 \\ -1 & \tau E \end{vmatrix} = 0 = s\tau D^2 + \tau D + 1,$$

$$D = -\frac{1}{2s} + i\sqrt{\frac{1}{s\tau} - \frac{1}{4s^2}}.$$

If, however, we replace D by $\Delta = e^D - 1$ we have

$$\begin{vmatrix} s(e^D - 1) + 1 & 1 \\ -1 & \tau(e^D - 1) \end{vmatrix} = 0 = s\tau e^{2D} - (2s\tau - \tau)e^D + s\tau - \tau + 1,$$

of which the solution is

$$D = \frac{1}{2}\log_e\left(1 - \frac{1}{s} + \frac{1}{s\tau}\right) + i\tan^{-1}\frac{\sqrt{s/\tau - {}^1/_4}}{s - {}^1/_2}.$$

If $s = 68.2$ and $\tau = 2$ the values of the pure imaginary parts of the two expressions for D will be nearly the same and thus the periods will also be; but the real part in the second case will be about $-{}^1/_{4s}$ whereas in the first it is $-{}^1/_{2s}$, i.e., twice as much.[7] This explains why Soper found numerically about 0.80 instead of about 0.58.

We have therefore seen that the damping on the "instant hypothesis" is determined (approximately) by the logarithmic decrement $Ap/4m$ instead of zero as stated by Soper and by $Ap/2m$ on the "hypothesis of distributed infectiousness." We have further seen that the damping is modified by replacing derivatives by increments. The rough equations given in an earlier paper[3] for the stepwise calculation of an epidemic with recruits, when applied to a variety of hypothetical cases, appear to give no damping and, as the theory indicates damping, those equations presumably have actually eliminated the lag to which damping would be due. However, it must be admitted that the phenomenon of recurrent measles epidemics gives no clear evidence of any damping. This creates something of a difficulty with the theory in respect to the prediction of damping and throws some doubt on the reality of periods; it is possible that measles simply dies out and then returns and under such a hypothesis there would seem to be no reason to expect either definite periods or damping to be observable by comparing successive epidemics.

[1] Soper, H. E., "The Interpretation of Periodicity in Disease Prevalence," *Jour. Roy. Statist. Soc., London*, 92, 34–61 (1929).

[2] If $\alpha = A\tau/m$ and the root for τD be $X + iY$ we have to solve $X - Xe^{-X}\cos Y - Ye^{-X}\sin Y + \alpha = 0$, $Y - Ye^{-X}\cos Y + Xe^{-X}\sin Y = 0$ and the solutions are of the form $X = -{}^1/_4\,\alpha\,[1 + c_1\alpha + c_2\alpha^2 + \ldots]$, $Y = \alpha^{1/2}\,[1 + d_1\alpha + d_2\alpha^2 + \ldots]$. The value of α for measles is a small number; if m be $5{}^1/_2$ years of recruits to the susceptible population and τ be half a month, $\alpha = {}^1/_{132}$.

[3] See these PROCEEDINGS, **28**, 361–367 (1942) and **31**, 24–34, 109–116 (1945).

[4] The value of aD may be written as a power series in pAa/m. If we set

$$R^2 = \frac{pAa}{m},\ L = \frac{1}{2} + \frac{\sigma^2}{24a^2},\ M = \frac{1}{6} + \frac{\sigma^2}{24a^2},\ N = \frac{1}{24} + \frac{\sigma^2}{48a^2} + \frac{\sigma^4}{1920a^4},$$

$$AD = iR\left[1 + \left(\frac{1}{2}M - \frac{5}{8}L^2\right)R^2\right] - \frac{1}{2}LR^2 + \left(\frac{1}{2}N - \frac{3}{2}LM + L^3\right)R^4$$

this second approximation will give an estimate of the accuracy of the first approximation when definite values of σ, τ, p, A, m are assumed. As $a = \tau + \sigma/2$ we must have σ/a varying from 0 to 2 as σ/τ varies from 0 to ∞. Soper really assumed $\sigma = 0$, $\tau = a$ fortnights for measles.

[5] The new-case-rate (of becoming infectious) at time t, viz., $A - dS/dt$, can be regarded as the sum of the products of S by the product

$$\frac{1}{m}\left(A - \frac{dS}{dt}\right)_\xi e^{(\xi - t)/\tau}\,d\xi = \frac{d\xi}{m}\left[\left(A - \frac{dS}{dt}\right)_\xi e^{(\xi - t)/\tau}\right] = \left(A - \frac{dS}{dt}\right)_\xi\left[\frac{1}{m}e^{(\xi - t)/\tau}\right]d\xi$$

contributed by the cases which become infectious at the time ξ previous to t, where the

contact rate is $r = 1/m$. The expression is capable of two interpretations, viz., (a) the contact rate remains constant but of the number of cases which become infectious at the time ξ only the fraction $e^{(\xi - t)/\tau}$ remain infectious, which seems to be Soper's hypothesis and which accords with the way in which the carrier condition in diphtheria falls off with the time, at least after the initial infectiveness (see Stallybrass, C. O., *Principles of Epidemiology*, pp. 302 and 617); (b) the cases which become infectious at time ξ all remain infectious but the contact rate r falls off as $e^{(\xi - t)/\tau}$. In the first interpretation τ is the average period of infectiousness, in the second it is the period required for the infectiousness to fall off to the fraction $1/e$ of its initial value; in the first we should speak of the diminishing incidence of the infectiousness, in the second of its diminishing intensity.

[6] See, for example, Wilson, E. B., *Advanced Calculus*, p. 451, Exs. 17, 18.

[7] If the root for D is $-X + iY$, the damping in time is e^{-Xt} and if X be divided by 2 the damping in the same time, being $e^{-Xt/2}$, is the square root of e^{-Xt}. There is, however, a further distinction which must be borne in mind. The differential equations are in terms of $u = \log (C/A)$. If C/A is near 1 so that the epidemic is infinitesimal, it is not the case-rate C which is directly subject to damping, but the part of it which exceeds A as $\log(C/A)$ is practically equal to $(C - A)/A$. If C is really near to A, the value of C is practically not damped appreciably. With $A = 2200$, $C_0 = 6600$, Soper's calculation gave him a second peak of $C_1 = 5168$ which is the fraction 0.783 of the first, and may be taken as a sort of damping factor—very likely the sort an epidemiologist would think most natural. However, to compare with the theory of the infinitesimal epidemic one would have to take $u_0 = \log (6600/2200) = 1.099$ and $u_1 = \log(5168/2200) = 0.8544$. The ratio of 0.8544 to 1.099 is 0.777 which is very near to 0.783. The theoretical value for the infinitesimal epidemic is 0.584 but when $\Delta = e^D - 1$ replaces D as an operator the theoretical solution gives, not 0.584, but 0.764 which certainly is as near as could be expected to the value 0.777 Soper found.

Of course, an epidemic with $C_0/A = 3$, $u_0 = 1.099$ can hardly be considered infinitesimal; the period will doubtless have increased somewhat (as Soper actually found in his calculation) and the theoretical damping will presumably have been somewhat modified. It should be noted that with $u_0 = 1.099$ we have, if we neglect effects of damping, from the equation $u_1 - e^{u_1} + e^{u_0} - u_0 = 0$ between the trough u_1 and crest u_0, the value $u_1 = -1.72$; the values $+1.10$ and -1.72 are so far from symmetrically situated with respect to $u = 0$ as to indicate a major departure from the simple harmonic deviation of the infinitesimal oscillation. Indeed it can no longer be expected that the damping as estimated by the fraction that one crest is of the preceding one and the fraction that one trough is of the preceding one should be the same. A satisfactory approximate solution of our equation (3) of a previous paper[8]

$$\left.\frac{du}{dt}\right|_{t+\tau/2} - \left.\frac{du}{dt}\right|_{t-\tau/2} = \frac{pA}{m} e^{u(t-\tau/2) - u(t+\tau/2)/p} \left(1 - e^{u(t+\tau/2)}\right)$$

would be required probably to discuss periodicity and damping in the case of finite epidemics.

TOPOLOGICAL METHODS IN THE THEORY OF FUNCTIONS OF A SINGLE COMPLEX VARIABLE. I. DEFORMATION TYPES OF LOCALLY SIMPLE PLANE CURVES*

BY MARSTON MORSE AND MAURICE HEINS

INSTITUTE FOR ADVANCED STUDY

Communicated July 25, 1945

1. Canonical Curves.—Title I refers to the first of a series of four papers by the authors under the general heading of "Topological Methods in the Theory of Functions of a Single Complex Variable." An abstract of Paper II will follow this abstract.

In the study of meromorphic functions one comes most naturally to *locally simple* sensed closed plane curves and deformations of these curves through families of sensed curves which are *uniformly* locally simple. A closed curve g is *locally simple* if for some sufficiently small positive constant e each point P of g is an interior point of a simple subarc of g whose end-points P_1 and P_2 satisfy the condition

$$|P_1P| > e \qquad |P_2P| > e.$$

A set of closed curves whose points P satisfy the preceding conditions for one and the same e will be termed *uniformly* locally simple. A curve may have infinitely many multiple points and be locally simple, and there are continuous families of locally simple closed curves which are not uniformly locally simple.

We admit deformations through continuous families of uniformly locally simple closed curves.

A first theorem is that any locally simple sensed closed plane curve g can be admissibly deformed into a curve C^n obtained by tracing a sensed circle

$$x = \cos\theta \qquad y = \sin\theta \tag{1.1}$$

n times, $n = \pm 1, \pm 2, \ldots$ or into a figure eight, C°. Moreover g can be deformed into but one of these canonical curves.

In particular the figure eight, C°, traced m times can be admissibly deformed into C°.

If a locally simple curve is given as the locally 1–1 continuous image of the circle (1.1) in the form

$$w = f(\theta) \qquad (w = u + iv)$$

then as θ increases from 0 to 2π any continuous branch of the function

$$\text{arc}\,[f(\theta + e) - f(\theta)] \qquad (0 < e \leq e_1)$$

will change by an integral multiple $2p\pi$ of 2π, independent of e provided e_1 is a sufficiently small positive constant. The integer p is called the *angular order* of g.

It is apparent that the angular order of an admissible curve g is invariant under any admissible deformation of g. It is shown that an admissible curve of angular order p can be admissibly deformed into the above canonical curve C^p, *including the figure eight* C^0 *when* $p = 0$.

A polygon π is locally simple if no edge has more than one end-point in common with the succeeding edge of the polygon. A continuous variation of a circular sequence of vertices representing a polygon will define an admissible deformation of π if successive vertices are bounded from each other and if successive edges have at most an end-point in common. We admit the addition or removal of a finite number of vertices on edges of a polygon at a finite number of moments t of a deformation.

The first approach to the preceding theorems is by way of lemmas by virtue of which any locally simple sensed curve can be admissibly deformed into an admissible polygon π. It is then shown that π can be admissibly polygonally deformed either into a triangle traced n times, $n = \pm 1, \pm 2, \ldots$, or into two triangles with a vertex in common forming a figure eight. The proof is by induction with respect to the number of multiple points.

2. *Products of Deformation Types.*—The class of admissible curves admissibly deformable into an admissible curve g will be denoted by (g) and termed a deformation type T. Each deformation type includes a sensed regular curve. With the aid of such regular curves the *product*

$$T_1T_2 = (g_1)(g_2)$$

will be defined. Let h_1 and h_2 be regular sensed curves in (g_1) and (g_2), respectively, which are positively tangent at a point P. Let h_1h_2 be a closed curve obtained by cutting h_1 and h_2 at P, to form arcs h' and h'', respectively, tracing h', then h'', and joining the final end-point of h'' to the initial end-point of h'. We define T_1T_2 as the deformation type (h_1h_2) and show that this type depends only on T_1 and T_2, and not on P or the choice of h_1 and h_2 in the types T_1 and T_2.

It is shown that the product T_1T_2 *is commutative and associative.*

For n positive or negative let g^n denote g traced n times. For $n = 0$ let g^n denote a curve of the type of $(g)(g^{-1})$. With these definitions it is shown that

$$(g^r)(g^m) = (g^{r+m})$$

where r and m are arbitrary integers. It must be clearly understood that the operation $(\quad)^m$ is *a priori* utterly different from the operation of tracing g m times.

It follows that the deformation types form an Abelian group with respect to the product operation, with the type of the figure eight the idem-factor. This group is isomorphic with the additive group of integers.

3. *The Order q.*—The order $q(w)$ of a sensed closed curve g with respect to point w not on g is well defined.

If g is locally simple, it is shown that the order $q(w)$ possesses an absolute bound independent of points w not on g.

An admissible deformation none of whose curves intersects a fixed point O will be called an *O-deformation*. It will be convenient to take O as the origin in the w-plane.

We need canonical curves with respect to O-deformations, given p and q. To that end let C be the sensed unit circle previously defined. Let C_1 be a similarly sensed circle of radius 1/4, tangent to C at $(u, v) = (1, 0)$. Let

$$C^q C_1^r \qquad (q \neq 0, r \neq 0) \tag{1.2}$$

denote the curve obtained by first cutting C and C_1 at $(1, 0)$, starting with the point $(1, 0)$, then tracing C q times in the sense indicated by the sign of q, next tracing C_1 similarly r times, closing the curve at $(1, 0)$. The curve C_1 shall be taken *interiorly* or *exteriorly* tangent to C at $(1, 0)$ according as q and r have the same or different signs. The resulting curve will be locally simple and regular. It is observed that the curve (1.2) has the order q with respect to $w = 0$ and the angular order p.

There exists an admissible curve g with prescribed order q and angular order p. A principal theorem follows:

Let g be an admissible curve of order q relative to the origin and of angular order p. Let the following cases be distinguished

Case I	$q \neq 0$	$p - q \neq 0$
Case II	$q \neq 0$	$p - q = 0$
Case III	$q = 0$.	

Then g admits an O-deformation into a curve $C^q C_1^{p-q}$ in Case I, a curve of the form C^p in Case II, and a curve of the form C_1^p in Case III. No one of these canonical curves with the orders (p, q) admits an O-deformation into a canonical curve with a different pair of orders.

The relation of this deformation theory to the theory of meromorphic functions will become clear in Paper II.

* Morse, M., and Heins, M., "Topological Methods in the Theory of Functions of a Single Complex Variable: II. Boundary Values and Integral Characteristics of Interior Transformations and Pseudo-harmonic Functions. III. Causal Isomorphisms in the Theory of Pseudo-harmonic Functions." Papers I and II are to appear in the *Annals of Mathematics*.

Morse, M., "The Topology of Pseudo-harmonic Functions." Completed but not yet submitted for publication. This paper obtains the basic topological relations A upon which Paper II is based. Paper III requires the use of infinite groups, and obtains relations A as a special case. Relations A can be obtained of themselves more simply and are so obtained in the present paper.

TOPOLOGICAL METHODS IN THE THEORY OF FUNCTIONS OF A COMPLEX VARIABLE. II. BOUNDARY VALUES AND INTEGRAL CHARACTERISTICS OF INTERIOR TRANSFORMATIONS AND PSEUDO-HARMONIC FUNCTIONS*

BY MARSTON MORSE AND MAURICE HEINS

INSTITUTE FOR ADVANCED STUDY

Communicated July 25, 1945

1. *Interior Transformations.*—The objective of the authors is to find out what properties of meromorphic functions are essentially topological in character, and by the introduction of appropriate topological concepts and methods to extend and simplify the present theory. To that end it is natural to replace the class of meromorphic functions by their topological generalizations, the interior transformations, studied by Stoilow,[1] Whyburn[2] and others, and the class of harmonic functions by a similar topological generalization termed by us pseudo-harmonic functions. The point of the present series of papers is by no means the generality obtained in this way but rather the insight and control of the theory, and the new problems and results which are brought out.

Let $w = F(t) \not\equiv$ const. be a transformation of a neighborhood N in the t-plane of a point t_0 into a neighborhood N_0 of w_0 (possibly the point at infinity) where $F(t)$ is analytic at t_0 or has a pole at t_0. Let $t = \varphi(z)$ map z_0 onto t_0 and N homeomorphically onto a neighborhood N_1 of z_0. Then

$$w = F(\varphi(z)) = f(z)$$

will be termed an *interior transformation* of N_1 into N_0. The point z_0 will be termed a zero or pole of $f(z)$ if t_0 is a zero or pole, respectively, of $F(t)$, and the order of z_0 as a zero or pole of $f(z)$ will be taken as the order of t_0 as a zero or pole of $F(t)$. If the inverse of $F(t)$ has a branch point of the mth order belonging to the pair (t_0, w_0) the inverse of $f(z)$ will be said to have a branch point of the mth order belonging to the pair (z_0, w_0).

We shall be concerned with a limited connected region G in the z-plane bounded by ν Jordan curves

$$(\beta) = (\beta_1, \ldots, \beta_\nu).$$

Without loss of generality these curves can be taken as circles. We shall *admit* transformations $w = f(z)$ of $\bar{G}$ (the closure of G) into the extended w-plane such that $w = f(z)$ is an interior transformation of some neighborhood of each point of G and is continuous on $\bar{G}$ at each point of (β). Such a function $f(z)$ will be called an interior transformation of G.

The "integral characteristics" of $f(z)$ are as follows: $n(0)$ = the number of zeros of $f(z)$ on G; $n(\infty)$ = the number of poles of $f(z)$ on G; μ = the

number of ramification points with antecedents on G of the inverse of $f(z)$. These points are counted with their multiplicities.

The *order* q of a sensed closed curve g in the w-plane with respect to $w = 0$ is well defined provided g does not pass through $w = 0$. If g is locally simple in the sense of Morse and Heins, I (these PROCEEDINGS, preceding article), the *angular order* p of g is well defined. A basic theorem follows.

I. *If $w = f(z)$ is meromorphic on G and continuous at each point of (β), if the images g_i of the positively sensed boundary curves β_i are locally simple and do not intersect the origin, then*

$$n(0) + n(\infty) - \mu \geqq 2 - \nu + \Sigma q - \Sigma p \tag{1.1}$$

summing over the boundary images g_i.

Examples exist in which the equality in (1.1) is excluded. Stoilow[1, b] has considered the case of one boundary curve whose image is a Jordan curve and has mistakenly affirmed that

$$n(0) + n(\infty) - \mu = 1.$$

For the Stoilow case $p = q = 1$, so that the right member of (1.1) is 1 but, as stated, the inequality need not hold.

II. *If, however, one adds the hypothesis that the transformation $w = f(z)$ is 1–1 in some neighborhood relative to G of each boundary point of G, then (1.1) becomes an equality.*

The theorem with the equality has been proved by us for the case of interior transformations, the relation (1.1) in I only for meromorphic function. The proof of I depends upon the lemma that the antecedents of points of ramification cluster at no boundary curve β_i whose image g_i is locally simple, and this has been proved only when $w = f(z)$ is meromorphic.

A generalization of a theorem of Radó[3] is given among various applications.

2. *Pseudo-harmonic Functions.*—If $f(z)$ is an interior transformation of a neighborhood N of z_0 without pole at z_0, then $Rf(z) = U(x, y)$ will be said to be *pseudo-harmonic* on N. In case $f(z)$ has a zero or pole at z_0 but is not identically zero, then

$$\log|f(z)| = U(x, y) \tag{2.1}$$

will be said to be pseudo-harmonic on N except for a *logarithmic pole.* Functions $U(x, y)$ will be admitted on $\bar{G}$ which are pseudo-harmonic on some neighborhood of each point of G except for logarithmic poles and which are continuous on $\bar{G}$ at points of (β).

Saddle points of $U(x, y)$ on G are similar to saddle points of harmonic functions. To treat the boundary values we begin with *Boundary Condition A* by virtue of which the function $U^{(\beta)}$ defined by U on (β) has at most

a finite number of points of extremum. It is seen that the locus $U = c$ at a given level c consists of a finite number of closed arcs, or arcs which tend to definite points of (β). Neighboring an interior point or a boundary point which is not isolated at the level c, the regions "below c" alternate with the regions "above c" in a well-defined manner. The *multiplicity* μ of a saddle point P is one less than the number of these "sectors" below c neighboring P. The interior or saddle points of positive multiplicity together with the points of relative minimum of U are called the *critical points* of U.

Set M = the number of logarithmic poles of U, S = the number of saddle points of U on G, s = the number of boundary saddle points, m = the number of points of relative minimum of U, counting the saddle points with their multiplicities. Then

$$M - S = 2 - \nu + s - m. \tag{2.2}$$

It is noted that $U(x, y)$ may be without partial derivatives both on G and $\bar{G}$, and that if these partial derivatives exist they may be null on (β). Under such conditions the classical alternatives to our evaluations, such as the Kronecker integral index, are undefined and therefore inapplicable.

Boundary Conditions B.—There is a special evaluation of $s - m$ on the right of (2.2) when U is of class C' on a neighborhood of (β) and has a non-null gradient λ on (β). These conditions are termed *Boundary Conditions B*, and do not imply that U is differentiable in general on G.

To proceed one covers the points of (β) at which λ is normal to (β) and entrant into G, by a finite set Ω of non-intersecting open arcs which exclude from their closures each point of (β) at which λ is normal to (β) and exeunt from G.

Let $\lambda(P)$ be the projection of λ onto the tangent to (β) at the point P on (β).

The Vector Index J of (β).—Let h be any arc of Ω. If h is a curve β_i an index 0 is assigned to h. If h is an arc of Ω not a curve β_i, an index 1 is assigned to h. To each end-point P of h at which $\lambda(P)$ is entrant relative to h an index 1 will be assigned. To other end-points 0 will be assigned.

The *vector index* J_i of U on β_i is defined as $E_0 - E_1$, where E_0 is the sum of the indices of end-points of arcs of Ω and E_1 is the sum of indices of the arcs of Ω. To (β) will be assigned the vector index

$$J = \Sigma J_i.$$

A first theorem is that J is independent of the choice of the covering Ω among admissible coverings. Then one readily proves that

$$M - S = 2 - \nu + J$$

for the most general pseudo-harmonic function satisfying Boundary Conditions B. In the special case in which Boundary Conditions A and B both hold, J equals the number of points of maximum of $U^{(\theta)}$ at which λ is entrant minus the number of minimum of $U^{(\theta)}$ at which λ is entrant.

3. *Comparison Theorems for Harmonic Functions.*—We begin with the case where $U(x, y)$ is harmonic on the unit circle $x^2 + y^2 \leqq 1$ and continuous on the boundary. If $u(r, \theta)$ represents U in polar coordinates, it is found that when the boundary values $u(1, \theta) = p(\theta)$ are of class C' and $p'(\theta)$ satisfies a Hölder condition, then $u_r(1, \theta)$ exists, and

$$u_r(1, \theta_0) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{p(\theta_0 + \alpha) - p(\alpha)}{\cos\alpha}\right]d\alpha \qquad [\text{if } p'(\theta_0) = 0].$$

A boundary point θ is called *entrant* if $u_r(1, \theta) < 0$ otherwise *exeunt*. The importance of these distinctions has already been indicated.

A function $p(\theta)$ of the above sort will be termed *R-admissible* if it has at most a finite number of relative maxima and minima occurring at isolated points between which $p'(\theta)$ is never 0.

If $p(\theta)$ is R-admissible, has N absolute minima and N absolute maxima and no other extrema, then the harmonic function with boundary values $p(\theta)$ has just $N - 1$ saddle points on the region $r < 1$.

In any case the number S of saddle points on $r < 1$ is at most $N - 1$. If $p(\theta)$ is R-admissible and has $2N$ points of extremum and if the corresponding function $u(r, \theta)$ has $S < N - 1$ saddle points on $r < 1$, and no differential critical points on $r = 1$, it is possible to successively modify $p(\theta)$ on arbitrarily small neighborhoods of the extremum points so that $p(\theta)$ remains R-admissible and none of its extremum values and points are changed, while the number of saddle points of $u(r, \theta)$ on the region $r < 1$ assumes each integer from S to $N - 1$, inclusive.

If $p(\theta)$ is R-admissible, it is possible to admissibly modify $p(\theta)$ on an arbitrarily small neighborhood of any one of its points of relative minimum (maximum) so that the resulting harmonic function has no critical points on the region $r < 1$.

Various other comparison theorems are given for harmonic functions with and without logarithmic poles.

* Morse, M., and Heins, M., "Topological Methods in the Theory of Functions of a Single Complex Variable: I. Deformation Types of Locally Simple Plane Curves. II. The present paper. III. Causal Isomorphisms in the Theory of Pseudo-harmonic Functions." Papers I and II are to appear in the *Annals of Mathematics*. For an abstract of Paper I see these PROCEEDINGS, **31**, 299–301 (1945).

Morse, M., "The Topology of Pseudo-harmonic Functions." Not yet submitted for publication.

[1] Stoilow, S., (*a*) "Du caractère topologique d'un théorème sur les fonctions méromorphes," *C. R. de l'Acad. des Sci. de Paris*, **190**, 251–253 (1930). (*b*) *Leçons sur les principes topologiques de la théorie des fonctions analytiques*, Paris, 1938.

[2] Whyburn, G. T., *Analytic Topology*, New York, 1942.

[3] Radó, T., "Zur Theorie der mehrdeutigen konformen Abbildungen," *Acta Szeged*, 1, 55–64 (1922).

ABSOLUTE SCALAR INVARIANTS AND THE ISOMETRIC CORRESPONDENCE OF RIEMANN SPACES

BY T. Y. THOMAS

DEPARTMENT OF MATHEMATICS, INDIANA UNIVERSITY

Communicated August 13, 1945

Let R_n be a Riemann space (defined by a positive definite quadratic differential form). Denote by $I_1, \ldots, I_p$ with $1 \leq p \leq n$ a set of absolute scalar differential invariants of R_n. These invariants determine absolute scalar functions $I_1(x), \ldots, I_p(x)$ of the coordinates $x^1, \ldots, x^n$ of R_n with functional matrix

$$\left\| \begin{matrix} \dfrac{\partial I_1}{\partial x^1} & \cdots & \dfrac{\partial I_1}{\partial x^n} \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \dfrac{\partial I_p}{\partial x^1} & \cdots & \dfrac{\partial I_p}{\partial x^n} \end{matrix} \right\|.$$

If this matrix has rank p at all points of R_n the scalars $I_1, \ldots, I_p$ are functionally independent, and if the functional matrix of every $q > p$ scalar invariants has rank not exceeding p the above scalars $I_1, \ldots, I_p$ (whose functional matrix does not vanish over R_n) are said to constitute a fundamental set of scalar invariants of R_n. In the following we deal with conditions for the isometric correspondence of two Riemann spaces on the basis of their absolute scalar invariants. The discussion is of a local nature so that the term Riemann space as here used is identical with the concept of the neighborhood of a point of a Riemann space when the *im Grossen* viewpoint is adopted. The method depends primarily on the functional independence of scalar invariants rather than on the order of the differential invariants from which these scalar functions are derived. Differentiability requirements on the components $g_{\alpha\beta}(x)$ of the fundamental tensor of R_n are immediately obvious from the differential invariants employed in any instance and will not be stated explicitly.

A space R_n having a fundamental set of invariants $I_1, \ldots, I_p$ with $p \geq 1$ will be said to be of category p. Obviously p cannot exceed n. In particular if all scalar differential invariants of R_n reduce to constants as functions of the coordinates we said that R_n is of category 0.

If R_n is of category p, where $0 \leqq p \leqq n$, and is in isometric correspondence with an $\bar{R}_n$, then $\bar{R}_n$ is of category p. For if $p = 0$, R_n is a space of constant curvature. Hence $\bar{R}_n$ must be of constant curvature and from this fact it follows readily that all scalar functions $\bar{I}(\bar{x})$ of $\bar{R}_n$ are constants. Hence $\bar{R}_n$ is of category 0. In general the $p(\geqq 1)$ independent scalars $I_1, \ldots, I_p$ of R_n will be carried into p independent scalars $\bar{I}_1, \ldots, \bar{I}_p$ of $\bar{R}_n$ by the correspondence; hence $\bar{R}_n$ cannot be of category $q < p$. Similarly if there were $q > p$ independent scalars in $\bar{R}_n$ these w uld yield $q > p$ independent scalars in R_n contrary to the hypothesis that R_n is of category p.

Let $I_1, \ldots, I_n$ be a fundamental set of scalar invariants of a space R_n of category n. For k fixed, the quantities $I_{k\alpha}$ (partial differentiation with respect to x^α) will be the components of a vector invariant of the space and hence the quantities $I_{jk} = g^{\alpha\beta} I_{j\alpha} I_{k\beta}$ will be absolute scalar invariants. Denoting by $\bar{I}_k$ and $\bar{I}_{jk}$ the corresponding scalars for a space $\bar{R}_n$ a necessary condition for the isometric correspondence of R_n and $\bar{R}_n$ is the consistency of the scalar equations

$$\bar{I}_j(\bar{x}) = I_j(x), \bar{I}_{jk}(\bar{x}) = I_{jk}(x) \tag{1}$$

where $j, k = 1, \ldots, n$. Let $x^\alpha = x^\alpha(\bar{x}^1, \ldots, \bar{x}^n)$ be a solution of (1). Then the $x^\alpha(\bar{x})$ will be differentiable functions; this follows from the theorem on the solution of implicit equations and the fact that the determinant $|I_{k\alpha}| \neq 0$ over R_n. Now from the first set of equations (1) we have

$$\bar{I}_{j\alpha}(\bar{x}) = I_{j\beta}(x) \frac{\partial x^\beta}{\partial \bar{x}^\alpha}.$$

From these relations and the second set of equations (1) we find

$$I_{j\mu} I_{k\nu} \left[\bar{g}^{\alpha\beta} \frac{\partial x^\mu}{\partial \bar{x}^\alpha} \frac{\partial x^\nu}{\partial \bar{x}^\beta} - g^{\mu\nu} \right] = 0.$$

Since $|I_{k\alpha}| \neq 0$ it follows that the bracket expression in the above equations must vanish. The solution $x^\alpha = x^\alpha(\bar{x})$ of (1) thus yields a (non-singular) isometric correspondence between R_n and $\bar{R}_n$. *Hence a space R_n of category n having a fundamental set of scalar invariants $I_1, \ldots, I_n$ can be put into isometric correspondence with a space $\bar{R}_n$ if, and only if, the equations (1) relating the coordinates x^α and $\bar{x}^\alpha$ of these spaces, admit a solution $x^\alpha = x^\alpha(\bar{x})$. If this solution exists it automatically gives the isometric correspondence.*

In what follows we limit our attention to spaces R_2. If R_2 is of category 2 the above result shows that five absolute scalars of R_2 suffice for the solution of the problem of the isometric correspondence between R_2 and any other space $\bar{R}_2$. Now suppose that R_2 is of category 1 and let J be the single independent scalar invariant of R_2, i.e., the matrix $\|\partial J/\partial x^\alpha\|$ has

rank 1 over R_2 (in particular J may be the Gaussian curvature K of R_2). Then form the scalars $J_1 = g^{\alpha\beta}J_\alpha J_\beta$ and $J_2 = g^{\alpha\beta}J_{\alpha\beta}$ where the $J_{\alpha\beta}$ are the components of the second covariant derivative of J.

LEMMA. *A space R_2 of category 1 having the fundamental scalar invariant J admits coordinates x, y (covering any point P of R_2) for which $J = x$ and the line-element has the form*

$$ds^2 = p(x)dx^2 + q(x)dy^2. \tag{2}$$

We can suppose $\partial J/\partial x^1 \neq 0$ at any point P of R_2 where x^1, x^2 are the coordinates of a suitably chosen system. Then $\eta^1 = J(x^1, x^2)$, $\eta^2 = x^2$ defines a non-singular transformation of a neighborhood of P. Let $g_{\alpha\beta} \rightarrow \theta_{\alpha\beta}(\eta)$ by this transformation. Now consider the equation

$$\theta^{1\beta}\frac{\partial v}{\partial \eta^\beta} = \theta^{11}\frac{\partial v}{\partial \eta^1} + \theta^{12}\frac{\partial v}{\partial \eta^2} = 0. \tag{3}$$

Since $\theta^{11} \neq 0$ this equation admits a solution $v(\eta^1, \eta^2)$ defined in the neighborhood of P, such that $\partial v/\partial \eta^2 \neq 0$ (see the discussion in C. Caratheodory, *Variationsrechnung und partielle Differentialgleichungen erster Ordnung*, 1935, p. 24). Hence the transformation $u = \eta^1$, $v = v(\eta^1, \eta^2)$ is non-singular and if $\theta_{\alpha\beta} \rightarrow k_{\alpha\beta}(u, v)$ relative to the u, v coordinates, then $k^{12} = 0$ by (3) where the indices 1 and 2 refer to the variables u and v, respectively. Hence $k_{12} = 0$. Also $J = u$ and hence $J_1 = k^{11} = f(u)$ since J_1 is functionally dependent on J by hypothesis. Since $k^{11} = 1/k_{11}$ it follows that $k_{11} = p(u)$. We now have

$$J_2 = -J_\sigma\Gamma^\sigma_{\alpha\beta}k^{\alpha\beta} = -k^{\alpha\beta}\Gamma^1_{\alpha\beta} = -k^{11}\Gamma^1_{11} - k^{22}\Gamma^1_{22}, \tag{4}$$

where we have used the Γ's to denote the Christoffel symbols relative to the u, v system.

Now $k^{\alpha\beta}{}_{,\gamma} = 0$. Taking $\alpha = \beta = \gamma = 1$ the resulting equation shows that Γ^1_{11} is a function of u alone. Since J_2 is a function of u alone, it now follows that the quantity $k^{22}\Gamma^1_{22}$ in the right member of (4) is a function of u alone, i.e.,

$$k^{22}\Gamma^1_{22} = -\frac{1}{2}\phi(u), \tag{5}$$

or,

$$\frac{\partial \log k_{22}}{\partial u} = k_{11}\phi(u) = \mathrm{p}(u)\phi(u) = \psi(u),$$

when we substitute the expression defining the Christoffel symbol Γ^1_{22} and replace the $k^{\alpha\beta}$ by their values in terms of the $k_{\alpha\beta}$ in (5). Integrating the

latter equation we see that h_{22} has the form $h_{22} = q(u)r(v)$. Then making the transformation

$$x = u, \; y = \int \sqrt{r(v)}\, dv$$

we obtain the form of the line-element given by (2), and $J = x$. This completes the proof of the lemma.

Now consider another space $\bar{R}_2$ of category 1. A necessary condition for the isometric correspondence of R_2 and $\bar{R}_2$ is that the three equations $\bar{J} = J$, $\bar{J}_1 = J_1$ and $\bar{J}_2 = J_2$ admit a solution $\bar{x}^\alpha = \bar{x}^\alpha(x^1, x^2)$ where $\bar{J}$, $\bar{J}_1$ and $\bar{J}_2$ denote the scalars in $\bar{R}_2$ which correspond to the scalars J, J_1 and J_2 in R_2. We assume the existence of this solution. Since $J_1 \neq 0$ in R_2 it follows that $\bar{J}_1 \neq 0$ in $\bar{R}_2$. Hence $\bar{J}$ is a fundamental scalar invariant for $\bar{R}_2$. Now suppose $P \to \bar{P}$ by the correspondence $\bar{x}^\alpha = \bar{x}^\alpha(x^1, x^2)$. Introduce the coordinates x, y in the neighborhood of the point P as given by the above lemma for the space R_2 and the corresponding coördinates $\bar{x}$, $\bar{y}$ in the neighborhood of $\bar{P}$ for the space $\bar{R}_2$ relative to which we have

$$\bar{J} = \bar{x}, \; d\bar{s}^2 = \bar{p}(\bar{x})d\bar{x}^2 + \bar{q}(\bar{x})d\bar{y}^2.$$

The above solution $\bar{x}^\alpha = \bar{x}^\alpha(x^1, x^2)$ implies the existence of a solution $\bar{x} = \bar{x}(x, y)$, $\bar{y} = \bar{y}(x, y)$ of the equations $\bar{J} = J$, $\bar{J}_1 = J_1$ and $\bar{J}_2 = J_2$ relative to these new coördinates. But from $\bar{J} = J$ we have $\bar{x} = x$. Also the relation $\bar{J}_1 = J_1$ yields $\bar{g}^{11}(\bar{x}) = g^{11}(x)$ from which we deduce $\bar{g}_{11}(\bar{x}) = g_{11}(x)$ or $\bar{p}(\bar{x}) = p(x)$. Then from $\bar{J}_2 = J_2$ we find

$$\bar{g}^{11}\bar{\Gamma}^1_{11} + \bar{g}^{22}\bar{\Gamma}^1_{22} = g^{11}\Gamma^1_{11} + g^{22}\Gamma^1_{22}. \tag{6}$$

Since g^{11} depends on x alone it follows that Γ^1_{11} depends on x alone. Similarly for $\bar{\Gamma}^1_{11}$. Also from the relations $\bar{g}^{11} = g^{11}$, $\bar{g}_{11} = g_{11}$ and $\bar{x} = x$ we obtain $\bar{\Gamma}^1_{11}(x) = \Gamma^1_{11}(x)$. Hence (6) reduces to $\bar{g}^{22}\bar{\Gamma}^1_{22} = g^{22}\Gamma^1_{22}$. On simplifying, this latter relation becomes

$$\frac{1}{\bar{q}}\frac{d\bar{q}}{d\bar{x}} = \frac{1}{q}\frac{dq}{dx}.$$

Then using the fact that $\bar{x} = x$ and integrating we have $\bar{q} = C^2q$ where C is a constant.

From the relations $\bar{p}(\bar{x}) = p(x)$, $\bar{q}(\bar{x}) = C^2q(x)$ and $\bar{x} = x$ we see immediately that the transformation $x = \bar{x}$, $y = C\bar{y} + d$ where d is a suitable constant gives an isometric correspondence between the neighborhoods of P and $\bar{P}$. Understanding that the term Riemann space is identical with a suitably restricted neighborhood in accordance with the local viewpoint the result obtained can now be stated as follows. *A Riemann space R_2 of category 1 with fundamental scalar invariant J can be put in isometric correspondence with a Riemann space $\bar{R}_2$ if, and only if, $\bar{R}_2$ is of category 1 and*

the three scalar equations $\bar{J} = J$, $\bar{J}_1 = J_1$ *and* $\bar{J}_2 = J_2$ *admit a solution* $\bar{x}^\alpha = \bar{x}^\alpha(x^1, x^2)$ *relating the coördinates* x^α *of* R_2 *to the coördinates* $\bar{x}^\alpha$ *of* $\bar{R}_2$.

Since a Riemann space R_2 of category 0 has its Gaussian curvature $K =$ const. we can immediately state the following result. *A Riemann space* R_2 *of category 0 can be put in isometric correspondence with a Riemann space* $\bar{R}_2$ if, and only if, $\bar{R}_2$ *is of category 0 and* $\bar{K} = K$ *where* K *and* $\bar{K}$ *are the Gaussian curvatures of* R_2 *and* $\bar{R}_2$, *respectively.* It has already been observed that the categories of two Riemann spaces in isometric correspondence must be the same. The remainder of the proof consists in showing that two Riemann spaces of the same constant curvature can be put in isometric correspondence and this is a well-known result.

We leave open the problem of the extension of these results to Riemann spaces R_n where $n > 2$. Another problem, worthy of serious study, concerns the investigation of the isometric correspondence of Riemann spaces in the large (for example, open and compact spaces) on the basis of their absolute scalar invariants.

BERNOULLI'S NUMBERS AND CERTAIN ARITHMETIC QUOTIENT FUNCTIONS

BY H. S. VANDIVER

DEPARTMENT OF PURE MATHEMATICS, UNIVERSITY OF TEXAS

Communicated July 6, 1945

In another paper[1] the writer established a congruence involving Bernoulli numbers, which can be extended, as we shall now show, to the generalized Bernoulli numbers of the first order. Such a number, $b_n(g, h)$, is defined as the number obtained by expanding $(gb + h)^n$ in full by the binomial theorem and substituting b_k for b^k where b_k is obtained by the use of the formula $(b + 1)^i = b_a$, $i > 1$, and the left-hand member is expanded as before and b_k is substituted for b^k. Let p be an odd prime. Then $p - 1$ has the form mc, with c even. Now consider the expression

$$\sum_{\rho}(1 - (d\rho + 1)^{p-1})^2 \equiv N \pmod{p^2}, \tag{1}$$

where ρ ranges over all the values such that $\rho^c \equiv 1 \pmod{p^2}$. Any term in this expression reduces to unity, modulo p^2, if ρ is such that

$$d\rho + 1 \equiv 0 \pmod{p}. \tag{2}$$

For each ρ that does not satisfy (2), the corresponding term of (1) is divisible by p^2. In the relation $x^c \equiv 1 \pmod{p^2}$, the c incongruent solutions modulo p^2 are also incongruent modulo p. Hence, N in (1) is the number

of incongruent solutions ρ, modulo p, in (2) and hence equals 1 or 0, according as $d^c \equiv 1 \pmod p$ or $d^c \not\equiv 1 \pmod p$. Expansion of (1) gives

$$N \equiv c - 2c \sum_{k=0}^{m} d^{kc} \binom{p-1}{kc} + c \sum_{j=0}^{2m} d^{jc} \binom{2(p-1)}{jc}$$

modulo p^2, and this may be written

$$N \equiv -2c \sum_{k=1}^{m} d^{kc} \binom{p-1}{kc} + c \sum_{k=1}^{m} d^{kc} \binom{2(p-1)}{kc} + c \sum_{k=1}^{m} d^{(m+k)c} \binom{2(p-1)}{(m+k)c}.$$

Set $d = ga + h$ and multiply through by $(ga + h)^n$, with $n \not\equiv 0 \pmod c$. Using the value of N noted above, we have, after letting a range over the integers $0, 1, 2, \ldots, p - 1$ and adding,

$$\sum_r (gr + h)^n \equiv -2c \sum_{k=1}^{m} S_{kc+n}(g, h)\binom{p-1}{kc} + c \sum_{k=1}^{m} S_{kc+n}(g, h)\binom{2(p-1)}{kc} + c \sum_{k=1}^{m} S_{mc+kc+n}(g, h) \binom{2(p-1)}{mc+kc} \quad (3)$$

modulo p^2, where

$$S_i(g, h) = \sum_{a=0}^{p-1} (ga + h)^i$$

and r ranges over the integers in the set $0, 1, 2\ldots, p - 1$ with $(gr + h)^c \equiv 1 \pmod p$. It can be shown, as follows, that for i even and $p > 3$

$$S_i(g, h) \equiv pb_i(g, h) \pmod{p^2}. \quad (4)$$

As noted in another[2] (relation (6), page 576 of the first reference) paper,

$$g(i + 1)S_i(g, h) = \sum_{q=0}^{i} \binom{i+1}{q+1} b_{i-q}(g, h)p^{q+1}g^{q+1}$$

which gives

$$S_i(g, h) = \sum_{q=0}^{i} \binom{i}{q} b_{i-q}(g, h)p \frac{p^{q-2}}{q+1} p^2 g^q. \quad (5)$$

For $q > 1$

$$p^{q-1} > (1 + 2)^{q-1} \geq 1 + 2(q - 1) = 2q - 1 \geq q + 1.$$

Hence each term of (5), beyond the second, is divisible by p^2, since the factor p cannot occur in the denominator of any fractional part oftener than in the numerator. But the second term will also be divisible by p^2, for i is

even and $b_{i-1}(g, h)$ is an integer.[2] This leaves (4). For i to be even in (4), n must be even, and the limitation $p > 3$ will be met; because if $p = 3$, then, as $p - 1 = cm$, $c = 2$, but $n \not\equiv 0 \pmod{c}$. Using (4) and dividing (3) through by p, we obtain, modulo p, with c and n even,

$$\frac{\sum_r (gr + h)^n}{p} \equiv -2c \sum_{k=1}^{m} \binom{p-1}{kc} b_{kc+n}(g, h) +$$
$$c \sum_{k=1}^{m} \binom{2(p-1)}{kc} b_{kc+n}(g, h) + c \sum_{k=1}^{m} \binom{2(p-1)}{mc+kc} b_{mc+kc+n}(g, h). \quad (6)$$

Reducing the binomial coefficients, modulo p, as on page 58 of a previous paper,[1] we obtain modulo p,

$$\frac{\sum_r (gr + h)^n}{p} \equiv -2c \sum_{k=1}^{m} (-1)^{kc} b_{kc+n}(g, h) +$$
$$c \sum_{k=1}^{m} (-1)^{kc}(kc + 1) b_{kc+n}(g, h) + c \sum_{k=1}^{m} (-1)^{kc-1} kc\, b_{p-1+kc+n}(g, h). \quad (7)$$

Now use the relation[2]

$$\frac{b_{s+p-1}(g, h)}{s + p - 1} \equiv \frac{b_s(g, h)}{s} \pmod{p}$$

where $g \not\equiv 0 \pmod{p}$ and for $s \not\equiv 0 \pmod{\overline{p-1}}$, which latter limitation is here complied with since $n \not\equiv 0 \pmod{c}$. We may then write, modulo p, with $g \not\equiv 0 \pmod{p}$,

$$\frac{\sum_r (gr + h)^n}{p} \equiv c \sum_{k=1}^{m} (kc + 1) b_{kc+n}(g, h) -$$
$$c \sum_{k=1}^{m} kc \frac{kc + n - 1}{kc + n} b_{kc+n}(g, h) - 2c \sum_{k=1}^{m} b_{kc+n}(g, h) \equiv -cn \sum_{k=1}^{m} \frac{b_{kc+n}(g, h)}{kc + n}.$$

This gives

THEOREM I. *If p is an odd prime with $p - 1 = mc$, c even, and $g \not\equiv 0 \pmod{p}$, then*

$$\frac{\sum_r (gr + h)^n}{p} \equiv -cn \sum_{k=0}^{m-1} \frac{b_{kc+n}(g, h)}{kc + n} \pmod{p} \quad (8)$$

for n even and not divisible by c, where the summation on the left extends over all integers r in the set $0, 1, 2, \ldots, p - 1$, such that $(gr + h)^c \equiv 1 \pmod{p}$.

For the special case $g = 1$, $h = 0$ this is Theorem I of the previous[1] paper, and we shall now proceed to find another congruence involving $M(c, n, p)$, that symbol being defined as the left-hand member of (8) for this special

case. In another[4] paper the writer defined a repetitive set modulo l as a set of integers

$$r_1, r_2, \ldots, r_c \tag{9}$$

where l is any integer and $r_i < l$, such that there exists an $r \not\equiv 1$ with

$$rr_1, rr_2, \ldots, rr_c$$

being congruent in some order to (9) modulo l. Hence, for $l = p$, the integers r in the set $1, 2, \ldots, p-1$, which satisfy $r^c \equiv 1 \pmod{p}$, form a repetitive set modulo p with any such $r \not\equiv 1$ as multiplier. As the writer proved, the set of integers

$$\frac{y_a p + r_a}{r}; \ a = 1, 2, \ldots, c$$

where

$$y_a \equiv -\frac{r_a}{p} \pmod{r}; \ 0 \leq y_a < r$$

is a permutation of (9). Hence

$$\sum_a r^n \left(\frac{y_a p + r_a}{r}\right)^n \equiv \sum_a r_a^n + np \sum_a y_a r_a^{n-1} \pmod{p^2}$$

or

$$(r^n - 1)M(c, n, p) \equiv n \sum_a y_a r_a^{n-1} \pmod{p}. \tag{10}$$

If

$$y_a \equiv -\frac{a}{p} \pmod{r}, \ 0 \leq y_a < r,$$

then for the special case $c = p - 1$ (10) gives

$$n \sum_{a=1}^{p-1} y_a a^{n-1} \equiv (r^n - 1)\frac{S_n}{p} \pmod{p},$$

and it follows that

$$\sum_{a=1}^{p-1} y_a a^{n-1} \equiv \frac{(r^n - 1)b_n}{n} \pmod{p}, \tag{11}$$

for $n \not\equiv 0 \pmod{\overline{p-1}}$.

Let $n = cs + e$, $0 < e < c$, where s ranges over the values $0, 1, \ldots, m-1$, and let r belong to the exponent c modulo p. (11) gives on addition

$$(r^e - 1)\left(\sum_{s=0}^{m-1} \frac{b_{cs+e}}{cs+e}\right) \equiv \sum_s \sum_{a_1} y_{a_1} a_1^{cs+e-1} + \sum_s \sum_{a_2} y_{a_2} a_2^{cs+e-1} \pmod{p} \tag{12}$$

where a_1 ranges over the a's satisfying $a^c \equiv 1 \pmod{p}$ and a_2 over the a's satisfying $a^c \not\equiv 1 \pmod{p}$.

We have

$$\sum_s a_1^{cs+e-1} = a_1^{e-1}(1 + a_1^c + a_1^{2c} + \ldots + a_1^{(m-1)c}) \equiv a_1^{e-1} m \pmod{p},$$

and

$$\sum_s a_2^{cs+e-1} = a_2^{e-1}(1 + a_2^c + a_2^{2c} + \ldots + a_2^{(m-1)c}) \equiv a_2^{e-1}\frac{a_2^{p-1} - 1}{a_2^c - 1} \equiv 0 \pmod{p}$$

since $a_2^c \not\equiv 1$. Hence, the right hand member of (12) reduces modulo p to

$$\sum_{a_1} m y_{a_1} a_1^{e-1}$$

Multiplication by $-ce$ gives

$$-ce(r^e - 1)\left(\sum_{s=0}^{m-1} \frac{b_{cs+e}}{cs+e}\right) \equiv \sum_{a_1} e y_{a_1} a_1^{e-1}. \tag{13}$$

But we also have (10). Since r belongs to the exponent c modulo p, and $0 < e < c$, and

$$\frac{b_{s+p-1}}{s+p-1} \equiv \frac{b_s}{s} \pmod{p}$$

for $s \not\equiv 0 \pmod{\overline{p-1}}$, comparison of (13) and (10) also yields the above noted special case of Theorem I for $g = 1$, $h = 0$.

It is not clear how the argument used in this second proof of Theorem I of the former[1] paper can be extended to give a proof of Theorem I of the present paper, since if we attempt this extension it seems necessary to employ a relation in another paper[4] (last separate congruence on page 123) and that relation was proved under a number of restrictions which would not apply in the statement of our general theorem.

[1] These PROCEEDINGS, **31**, 55–60 (1945), Theorem I. On page 55 of this paper on the second line above (4), insert the factor $(-1)^n$ in the right-hand member. The relation (4) is subject to the condition $n \not\equiv k\mu$. On page 56, second line, read "are at most only $(p-3)/2$" in lieu of "are only $p - 1$." In the first line above relation (6), insert 0 after the second congruence sign. In the eighth line above relation (7), read $a < p$ in lieu of $a < (p-1)$. On page 59, fourth line above relation (14), read $\sum_r r^n$ in lieu of $\sum_r r$. In connection with the last formula on page 60, add the condition $(n(n + c), p) = 1$.

[2] These numbers were considered by the writer in *Duke Math. Jour.*, **8**, 575–584 (1941), and also in *Trans. Am. Math. Soc.*, **51**, 512–519 (1942).

[3] These PROCEEDINGS, **28**, 326 (1942), relation (8) for $j = 1$.

[4] *Bull. Am. Math. Soc.*, **46**, 122 (1940).

THE TOTAL DIFFERENTIAL EQUATION FOR THE EXPONENTIAL FUNCTION IN NON-COMMUTATIVE NORMED LINEAR RINGS

By Aristotle D. Michal

California Institute of Technology

Communicated June 30, 1945

A normed linear ring[1] is a normed linear space (say with real multipliers) over which there exists an associative (but not necessarily commutative) function AB (additive and continuous in each variable) with values in the space. We shall assume the existence of a unit element I and assume that the space is a complete normed linear space (Banach space). We need not normalize the modulus of AB and take it as the number 1, nor need we normalize the norm of the unit element I and take it as the number 1. Examples of normed linear rings are sets of square matrices, quaternions and more generally any (real) associative algebra topologized with a norm topology in a variety of possible ways. Important infinite dimensional examples abound in the theory of integral equations and functionals.

The exponential function e^A in a normed linear ring is defined by

$$e^A = I + A + \frac{A^2}{2!} + \ldots + \ldots.$$

Since $\|AB\| \leq m\|A\|\,\|B\|$, where m is the modulus of the bilinear function AB, it follows that the exponential function in a normed linear ring is an entire analytic function.[2] The object of this paper is to characterize the exponential function in a normed linear ring by a total differential system in Fréchet differentials. More specifically, this characterization is embodied in the following theorem.

THEOREM. *The non-linear total differential system in Fréchet differentials of functions $z(A)$ with arguments and values in a normed linear ring N*

$$\left.\begin{array}{l} \delta z(A) = \int_0^1 z((1-\xi)A)\delta A z(\xi A)d\xi \\ z(0) = I, \text{ the unit of } N, \end{array}\right\} \qquad (1)$$

has a unique entire analytic solution given by the exponential function $z(A) = e^A$ in N.

Let

$$z(A) = \sum_{i=0} z_i(A) (z_i(A), \text{ a homogeneous polynomial of degree } i)$$

be an entire analytic solution of (1). Clearly $z_0(A) = I$ and by calculation $z_1(A) = A$, $z_2(A) = A^2/2!$. In this and in what follows, a theorem by the

author[3] on term by term Fréchet differentiation of a power series in normed linear spaces is used. The following recurrence formula holds for $n \geq 2$ and for all A, $\delta A \in N$:

$$(n+1)P_{n+1}(A, A, \ldots, A, \delta A) = \int_0^1 \{\delta A z_n(\xi A) + z_n((1-\xi)A)\delta A + \sum_{\substack{i+j=n \\ i,j \geq 1}} z_i((1-\xi)A)\delta A z_j(\xi A)\} d\xi, \quad (2)$$

where $P_{n+1}(A_1, A_2, \ldots, A_{n+1})$ is the polar of the polynomial $z_{n+1}(A)$. The truth of the theorem follows after some calculation with the aid of the recurrence relation (2).

The ideas that led the author to the differential system (1) for the particular case in which the normed linear ring N is a matric ring are discussed in the author's paper, "Differential Equations in Fréchet Differentials Occurring in Integral Equations," these PROCEEDINGS, **31**, 252–258 (1945).

The characterization of other elementary functions in a normed linear ring N, such as sin A and cos A, can also be effected by total differential systems in Fréchet differentials. Similar results can also be given for normed linear rings without a unit. One can, for example, study the properties of the function defined by the expansion

$$A + \frac{A^2}{2!} + \frac{A^3}{3!} + \ldots + \ldots$$

It should be pointed out that the differential system (1) is not of the type for which Michal and Elconin have given existence and uniqueness theorems.[4]

[1] A search through the literature seems to reveal that the first systematic study of normed linear rings (non-Hilbert variety) was made by Martin, R. S., and Michal, A. D., in "Some Expansions in Vector Space," *Jour. Math. Pures et Appl.*, **13**, 69–91 (1934). This paper was presented before the Los Angeles 1932 national meeting of the American Mathematical Society. The author has made an extensive use of normed linear rings in his studies on general differential geometry. See Michal, A. D., "General Differential Geometries and Related Topics," *Bull. Amer. Math. Soc.*, **45**, 529–563 (1939), where earlier references are included. See also Michal, A. D., and Mewborn, A. B., "Abstract Flat Projective Differential Geometry," *Acta Mathematica*, **72**, 259–281 (1940). For problems in mathematical analysis, besides those treated in the Michal-Martin paper, the reader is referred to Michal, A. D., and Hyers, D. H., "Second Order Differential Equations with Two Point Boundary Conditions in General Analysis," *Amer. Jour. Math.*, **58**, 646–660 (1936); and to Michal, A. D., and Elconin, V., "Completely Integrable Differential Equations in Abstract Spaces," *Acta Mathematica*, **68**, 71–107 (1937). The last paper also considers briefly some questions on ideals in normed linear rings. In recent years many important contributions have been made to the theory of normed linear rings by many authors in Russia as well as in America. The Russian mathematician I. Gelfand is especially to be mentioned in this connection. See his paper in *Rec. Math.* (*Mat. Sbornik*), **51**, 3–24 (1941). For commutative normed linear rings, see Lorch, E. R., *Bull. Amer. Math. Soc.*, **50**, 447–463 (1944).

[2] Cf. Michal-Martin paper, loc. cit.

[3] Michal, A. D., *A Theorem on the Fréchet Differentials of Power Series in Normed Linear Spaces* (unpublished).

[4] Cf. Michal-Elconin paper, loc. cit.

UNIVERSAL RATIONAL FUNCTIONS

BY E. T. BELL

DEPARTMENT OF MATHEMATICS, CALIFORNIA INSTITUTE OF TECHNOLOGY

Communicated July 17, 1945

1. *Positive* in all that follows shall mean *greater than zero*. The coefficients in all rational functions considered are integers (actually all positive in the examples given).

According to a customary definition, a form F is universal if, for integer values (positive, zero or negative) of all the variables in F, F represents all integers, positive, zero or negative. Another type of universality, as in Waring's problem, requires that a form represent all positive or zero integers for positive or zero values of the variables. Thus the sum of four squares is universal in this sense, but is not universal if zero values of the variables are excluded. A further type restricts the universality to some proper subset C of all the integers, for example, cubes.

We shall call the rational function F/G, where F, G are forms with positive integer coefficients, and F/G is in its lowest terms, *universal*, if, for positive integer values of all the variables in F, G, F/G represents all positive integers. If F/G for positive integer values of the variables represents all integers in the subset C, F/G will be called *universal for C*. The universal rational functions considered here also represent all negative integers when negative integer values of the variables are admitted. These particular universal rational functions therefore represent all integers except zero for integer values of the variables different from zero. A point of interest, however, is that positive integer values of the variables suffice for the representation of all positive integers.

2. An example of a universal rational function in four variables x, y, z, w is

$$\frac{xz(x^{n-1}w^{n+1} + z^{n-1}y^{n+1})}{y^{n-1}x^{n+1} + w^{n-1}z^{n+1}},$$

where n is an arbitrary constant positive integer;

$$\frac{x^{m+1}y^{n+1}u^{s+1} + z^{m+1}w^{n+1}v^{s+1}}{x^{m-1}y^{n-1}u^{s-1} + z^{m-1}w^{n-1}v^{s-1}},$$

in six variables x, y, z, u, v, w, is universal for squares, where m, n, s are

arbitrary constant positive integers. In $2s$ variables $x_1, \ldots, x_s, y_1, \ldots, y_s$, with the m_i, n_i arbitrary constant positive integers and

$$F(x_1, \ldots, x_s, y_1, \ldots y_s) \equiv \prod_{i=1}^{s} x_i^{m_i+1} + \prod_{i=1}^{s} y_i^{n_i+1},$$

$$G(x_1, \ldots, x_s, y_1, \ldots, y_s) \equiv \prod_{i=1}^{s} x_i^{m_i-1} + \prod_{i=1}^{s} y_i^{n_i-1},$$

F/G is universal for squares. With $s > 1$ and

$$F \equiv x_s \prod_{i=1}^{s-1} x_i^{m_i+1} + y_s \prod_{i=1}^{s-1} y_i^{n_i+1},$$

$$G \equiv y_s \prod_{i=1}^{s-1} x_i^{m_i-1} + x_s \prod_{i=1}^{s-1} y_i^{n_i-1},$$

F/G is universal. For even integers,

$$\frac{x^2r^2 + y^2s^2 + z^2t^2 + 3w^2u^2}{xr + ys + zt}$$

is universal. For fourth powers, $(x^5 + u^2v^2w)/(x + w)$ is universal. Illustrative of an infinity of universal rational functions F/G in which each of F, G is a binomial, $x^2w^2(y^2 + zw)/(x^5 + w^5)$ is universal.

3. The foregoing examples suggest that universal rational functions are much more easily found or constructed than are universal forms. That this is so, is evident from the remark that an elementary monomial represents all positive integers for positive integer values of the variables: the monomial $x_1^{a_1} \ldots x_s^{a_s}$, in the s independent variables $x_1, \ldots, x_s$, is *elementary* if at least one of the positive integers $a_1, \ldots, a_s$ is 1. Without loss of generality, we shall assume $a_1 \leq a_2 \leq \ldots \leq a_s$; $(a_1, a_i, \ldots, a_s)$ is then the *index* of the monomial.

If F/G is universal, $F = nG$, where n is an arbitrary positive integer, has a solution for positive integer values of all the variables $y_1, \ldots, y_t$ in F, G. Suitable F, G are found by representing n as an elementary monomial, say $n = n_1^{a_1} \ldots n_s^{a_s}$, of index $(a_1, \ldots, a_s)$. For a given index, the total number of such representations is readily determined as a combinatorial function of the exponents in the prime decomposition of n. The representation of n in F/G is then of the type

$$y_i = M_i(n_i, \ldots, n_s), \; i = 1, \ldots, t,$$

where the M_i are monomials in $n_1, \ldots, n_s$.

Write $F/G \equiv R(y_1, \ldots, y_t)$. From what precedes, it follows that

$$R(z_1, \ldots, z_t)R(w_1, \ldots, w_t) = R(z_1w_1, \ldots, z_tw_t)$$

has an infinity of solutions in monomials $z_1, \ldots, z_t, w_1, \ldots, w_t$. For if

$n = n_1^{a_1} \ldots n_s^{a_s}$, $m = m_1^{a_1} \ldots m_s^{a_s}$ are representations of the arbitrary positive integers n, m in a monomial of index $(a_1, \ldots, a_s)$, their product is represented as $(n_1m_1)^{a_1} \ldots (n_sm_s)^{a_s}$ in a monomial of the same index. If $z_t = M_t(n_1, \ldots, n_s)$, $w_t = M_t(m_1, \ldots, m_s)$ represent n, m, respectively, in R,

$$z_t w_t = M(n_1m_1, \ldots, n_sm_s)$$

for the representation of nm in R.

It follows also that if $R(x_1, \ldots, x_t)$ is a universal rational function, and $R_1, \ldots, R_t$ are any universal rational functions in any variables, then $R(R_1, \ldots, R_t)$ is a universal rational function. For each variable denotes a positive integer, and hence each is representable in any universal rational function.

Appropriate F, G may be constructed indefinitely from the disjunction (logical product) and conjunction (logical sum) of simple multiplicative diophantine equations. If the totally distributed form of the disjunction or conjunction can be written as $F = MG$ (or as $G = MF$), where M is an elementary monomial, and M, G (or M, F) have no common factor, then F/G (or G/F) is universal. The solution of $F/G = n$ (or of $G/F = n$), where n is an arbitrary positive integer, is then obtained by expressing n in the form M. If $F = MG$ (or $G = MF$) was obtained by disjunction, the values of the variables are found by solving any one of the equations in the disjunction; if the equation was obtained by conjunction, the equations are solved simultaneously. Since in either case the solutions are the general solutions, given in a necessary and sufficient number of independent integer parameters, the corresponding M represent all positive integers. The first example in the preceding section was constructed from the disjunction of $xy^n = zw^n$, $yz^n = wx^n$; the fifth, from the conjunction of $xr = ys = zt = wu$. Both may be considerably generalized.

PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 31 October 15, 1945 Number 10

STRUCTURE OF THE SALIVARY GLAND CHROMOSOMES OF DIPTERA

BY HANS RIS AND HELEN CROUSE

ROCKEFELLER INSTITUTE FOR MEDICAL RESEARCH, NEW YORK, AND DEPARTMENT OF ZOÖLOGY, UNIVERSITY OF PENNSYLVANIA*

Communicated September 4, 1945

The giant chromosomes of the Diptera constitute ideal material for studies on chromosome structure and chemistry and for cytogenetic analysis because of their tremendous size and obvious pattern of longitudinal differentiation. So far it has been impossible to take full advantage of them, however, because they have not been fully understood in terms of the structure of mitotic chromosomes.

In the literature we find three types of interpretation of the banded appearance of the giant chromosomes: (1) The chromosome is composed of several helically coiled threads. The gyres of this coil appear as bands.[1-3] (2) The chromosome is formed of a bundle of completely relaxed chromonemata which originated endomitotically; because of somatic synapsis homologous chromomeres join to form bands.[4-7] (3) The chromosome consists of a large number of chromonemata which are submicroscopic and therefore invisible. The visible structures do not correspond to chromomeres or chromonemata but originate in a different manner.[8]

None of these hypotheses is wholly satisfactory. The first one has been discredited because it is impossible to interpret the bands of the mature giant chromosomes as gyres of a simple large coil. On the other hand, many investigators have seen coils in the giant chromosomes, and bands and coils have even been found to occur simultaneously in the same nucleus.[1, 2, 9-11] A satisfactory interpretation of the giant chromosomes must take this evidence into account.

The second hypothesis has been adequately criticized by Metz. It cannot account for the vesiculated appearance which the giant chromosomes often show, nor for the length of the chromonemata unless some growth of

the chromosome is assumed in addition to uncoiling. This theory, nevertheless, is accepted today by most cytologists because it allows a uniform interpretation of chromosome structure based on the chromomere hypothesis. Recent work on plant and animal chromosomes, however, has shown that the chromonema is uniform and not a series of chromatic granules on an achromatic thread.[12-16] The chromomeres are misinterpretations of coiled structures or points of overlap of chromonemata. Therefore, since mitotic chromosomes are not composed of true chromomeres, it seems doubtful that bands can be interpreted as aggregates of chromomeres.

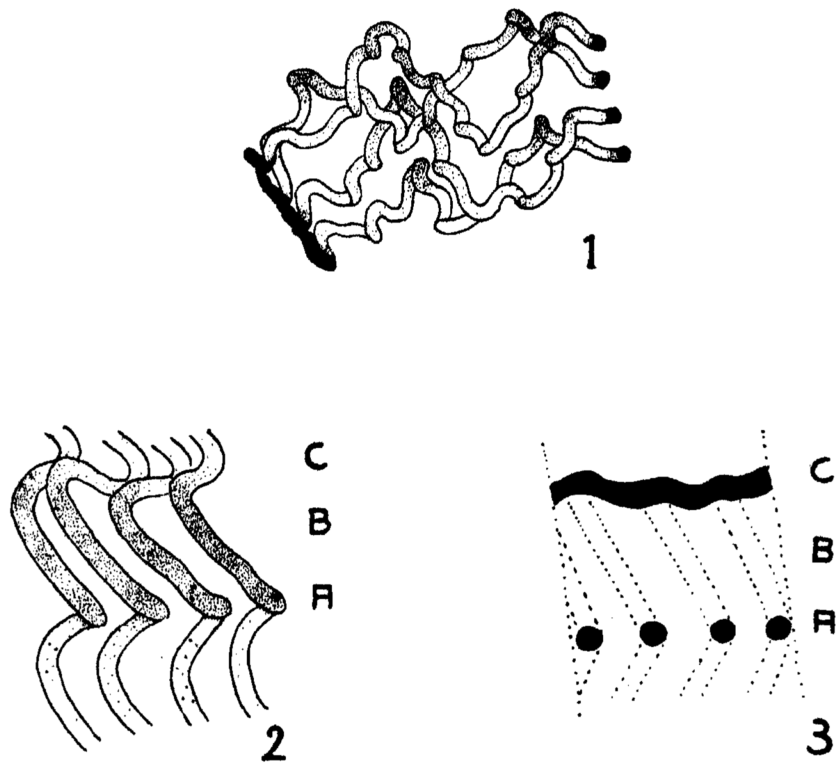
The third hypothesis, assuming submicroscopic chromonemata, places the giant chromosomes in a class by themselves since the component chromonemata are visible in all other types of chromosomes.

The purpose of this paper is to suggest an interpretation of the structure of giant chromosomes which is in harmony with the latest knowledge of chromosome structure in general (meiotic chromosomes of Tradescantia and the grasshopper, lamp-brush chromosomes of the frog oöcyte[15, 16]) and which can account for the apparently conflicting observations of different investigators. Proof of our hypothesis rests on a detailed study of the development of the giant chromosomes. This work is in progress.

According to our hypothesis the salivary chromosome (Sciara) consists of a definite number of chromonemata which are coiled in a complicated fashion. The coiled threads can be seen most clearly in regions where the banded organization has been disrupted (so-called puffed regions). These regions occur at definite loci along the chromosomes of the larva and become increasingly prominent in the pupa as histolysis sets in. Figure 1 illustrates our interpretation of such a puffed region. Along each chromonema the gyres of a narrowly pitched helix ("minor coil") can be seen. The chromonemata do not run parallel to the length of the chromosome but proceed in an irregular manner, weaving back and forth across the width of the chromosome. Chemical agents which are known to uncoil the mitotic chromosome (KCN, $NaHCO_3$, hot water) can transform normally banded material into a similar mass of threads. These observations suggest that the banded chromosome is composed of coiled chromonemata. It will now be demonstrated how the coiled chromonemata can form bands.

Chromosomes from untreated medium aged Sciara larvae are best suited for this demonstration. At this stage of development four chromonemata can be followed in each homologue. Figure 2 represents our interpretation of two successive bands in terms of coiled chromonemata. It can be seen that the chromonemata are thrown into wide gyres ("major coil") as they proceed along the length of the chromosome. In region *A* they are running vertically, and therefore in optical cross section give the appearance of a granular band (Fig. 3). From here each chromonema can be traced as it

runs horizontally across an interband region *B*. Because of the major coil, the chromonemata of the interband region always run diagonally as most investigators have observed. Only in greatly stretched areas do they run parallel to the long axis of the chromosome. The much lighter appearance of the interband regions we explain as follows: (1) the chromo-



FIGURES 1-3

Diagrammatic representation of the structure of salivary gland chromosomes in a medium-aged larva (Sciara). Only one homologue is drawn.

Fig. 1. Puffed region with bands disrupted.

Fig. 2. The course of the chromonemata through two consecutive bands. (The minor coil is omitted.)

Fig. 3. Appearance of the same region at a medium focal level.

nemata run vertically in the band regions and horizontally in the interbands; consequently, much more light is absorbed in the bands; (2) one usually focuses on a longitudinal optical section of the chromosome, for here the bands are clearest. In such a section, however, more chromo-

nemata are in focus in the band than interband regions; (3) the salivary chromosomes are usually examined in smear preparations. Smearing stretches the chromonemata disproportionately in the interband regions, for the bands, where the chromonemata run transverse to the chromosome axis, resist stretching. In sectioned material stained by the Feulgen reaction we find that chromonemata are much more conspicuous in the interbands. Region *C* in figure 2 illustrates how a solid (i.e., non-granular) band could arise. Here we follow the chromonemata into the next gyre of the major coil. As they dip down, running from left to right across the chromosome, the impression of a continuous line rather than separate granules is produced (Fig. 3). The wavy border of such bands is caused by the minor coil of the chromonemata. Because the individual threads of region *C* proceed at different levels, filling out most of the cross section of the chromosome, this region looks like a solid disc (i.e., threads at all levels). Sometimes in chromosomes of medium aged larvae several gyres of a two-stranded helix can be seen within one homologue.[1–3, 9–11] This appearance can best be understood if we assume that the four chromonemata are closely appressed in pairs, forming the wide gyres of the major coil. When these threads have separated laterally, a banded structure is produced as has been shown above. Such an interpretation makes it possible to understand how different regions on the same chromosome may appear as gyres of a helix or as typical bands.

Several authors have observed that the bands often look like rings in cross section through the chromosome instead of solid discs. Our hypothesis could account for the apparently contradictory observations. When the chromonemata are closely appressed and proceed in a common helix, the gyres of the major coil appear in optical cross section as rings. When the chromonemata have come apart, however, the gyres of the major coil fill the entire cross section of the chromosome, and the bands now look like solid discs.

In old larvae the banded chromosomes have the same structural characteristics, the only difference apparently being the greater number of threads brought about by endomitosis.

If the chromonemata in these giant chromosomes have both minor and major coils, as appears to be the case, they must have grown enormously in length. Such a growth of chromonemata is, however, not restricted to the banded chromosomes. A comparable increase in length occurs, for instance, in the "lamp-brush" chromosomes of certain oöcytes where recent studies[16] have shown that the characteristic loops are actually the gyres of the major coil of the chromonema and contain in addition the minor coil. In contrast to this great increase in length, the chromonema seems to remain approximately constant in diameter. Growth is thus restricted to the long axis of the chromosome. Some authors have suggested that the

longitudinal growth occurs mainly in "nongenic" material of the chromonema. Since giant chromosomes are always found in cells which are especially active physiologically, and since no visible differentiation into alternate "genic" and "inert" regions is visible, it seems to us more likely that the genes themselves increase in mass. Heterochromatic regions which have been shown to contain few genes (Drosophila) do not exhibit in the salivary chromosomes longitudinal growth comparable to that of euchromatic regions. This is in agreement with the view that the genes themselves grow rather than the inert, non-genic material.

The extensive cytogenetic work on Drosophila has shown that the bands in the giant chromosomes can be correlated with definite gene loci. If the bands are due to specific coiling, an interesting relationship between the coiling and the specific molecular structure of the chromonema becomes apparent. This complex coiling pattern of the giant chromosomes could be an expression of the longitudinal differentiation in the gene string. The detailed correspondence of coiling and gene specificity poses an interesting problem for investigation.

Cooper[17] has pointed out that the increase in number of strands in a helix causes it to uncoil. This, however, is true only as long as the component threads remain closely appressed in a common helix. It no longer holds if the chromonemata can separate laterally, as appears to be the case in the giant chromosomes.

It is generally believed that somatic as well as meiotic synapsis is the consequence of a complete uncoiling of the chromosomes. If this were true, it would constitute a serious objection to our assumption that the somatically synapsed giant chromosomes consist of coiled chromonemata. However, the demonstration of a typical helix in synapsing meiotic chromosomes of Tradescantia and the grasshopper[15, 16] invalidates this objection. The chromosomes are coiled when they synapse in meiosis, and they are coiled when they synapse somatically in young dipteran larvae.

It is generally assumed that the chromonemata within one chromosome are held together by the same forces that result in synapsis of homologues. This assumption is unjustified. Rather, it appears that the chromonemata within a chromosome are held together in the same fashion as the coiled chromatids of ordinary prophase chromosomes. It has been suggested[18] that achromatic material binds the chromatids together. Experiments with hypo- and hypertonic solutions on salivary chromosomes[19] seem to confirm this hypothesis. In hypotonic solutions the chromonemata separate, and bands disappear. This can be reversed in isotonic medium. With hypertonic Ringer the chromosome shrinks in diameter, and the bands appear more distinct. All these facts can best be explained by a reversible swelling and shrinkage of an achromatic substance between the coiled chromonemata.

The most puzzling appearance of the salivary chromosome is the so-called vesiculated condition. This can occur at localized regions in different developmental stages of the Sciara larva. In certain physiological states all the giant chromosomes of larvae in the same culture may be vesiculated. Under other conditions the banding pattern completely disappears and the chromosomes look like masses of faintly staining coiled threads (ghost chromosomes). Appearances similar to these can be induced experimentally in normally banded material. Immersion for a few seconds in 1 *M* NaCl causes the chromosomes to become vesiculated. Prolonged treatment with the salt solution produces typical ghost chromosomes. We interpret these phenomena in the following manner. Around each chromonema there is a Feulgen-positive substance, which, under certain conditions, can come off the threads and form chromatic connections between chromonemata and between chromosomes (chromatic coating[18]). Such connections between the coiled chromonemata of the giant chromosomes give the appearance of vesiculation. As was shown by Mirsky and Pollister,[20] nucleoproteins can be dissolved from chromosomes with 1 *M* NaCl. The vesicles observed in the giant chromosomes after short treatment with the salt solution we believe to be caused by the initial dissolution of this substance. Prolonged treatment completely dissolves these nucleoproteins; ghost chromosomes and a basophilic cytoplasm result. This basophily of the cytoplasm is likewise characteristic of untreated ghost cultures.

Summary.—On the interpretation presented here the giant chromosomes of dipteran larvae consist, like mitotic chromosomes, of a number of helically coiled chromonemata. The chromomeres in the giant chromosomes are misinterpretations of coiled structures, as has been demonstrated for meiotic and mitotic chromosomes. It is shown here how the appearance of bands and interbands may be caused by complex coiling of a bundle of chromonemata. The chromonema itself is uniformly Feulgen-positive. The giant size of these chromosomes would then be due to: (1) great increase in length of the chromonema (longitudinal growth of individual genes); (2) increase in the number of chromonemata by endomitosis; (3) lateral separation of the coiled chromonemata.

* Aided by a grant from the Rockefeller Foundation to Dr. C. W. Metz.

[1] v. Herwerden, M. A., *Anat. Anz.*, **36**, 193–207 (1910).
[2] Kaufmann, B. P., *Am. Nat.*, **65**, 555–558 (1931).
[3] Sinotô, Y., and Yuasa, A., *Jap. Jour. of Gen.*, **10**, 245–248 (1935).
[4] Koltzoff, N., *Science*, **80**, 312–313 (1934).
[5] Bridges, C. B., *Jour. of Hered.*, **26**, 60–64 (1935).
[6] Painter, T., *Am. Nat.*, **73**, 315–330 (1939).
[7] Bauer, H., *Zeit. Zellf.*, **23**, 280–313 (1935).
[8] Metz, C. W., *Cold Spring Harbor Symposia on Quantitative Biology*, **9**, 23–36 (1941).
[9] Alverdes, F., *Arch. Zellf.*, **9**, 168–204 (1912).

[10] Tänzer, E., *Zeit. wiss. Zool.*, **119**, 114–153 (1921).
[11] Dawydov, W., *Zeit. Zellf.*, **10**, 625–641 (1930).
[12] Kaufmann, B. P., *Am. Nat.*, **65**, 280–282 (1931).
[13] Koshy, T. K., *Jour. Roy. Microsc. Soc.*, **54**, 104–120 (1934).
[14] Naithani, S. P., *Ann. Bot. N. S.*, **1**, 257–276 (1937).
[15] Swanson, C. P., *Am. Jour. Bot.*, **30**, 422–428 (1943).
[16] Ris, H., (in press).
[17] Cooper, K. W., these PROCEEDINGS, **24**, 452–458 (1938).
[18] Ris, H., *Jour. Exp. Zool.*, **90**, 267–322 (1942).
[19] Doyle, W. L., and Metz, C. W., *Biol. Bull.*, **69**, 126–135 (1935).
[20] Mirsky, A. E., and Pollister, A. W., these PROCEEDINGS, **28**, 344–352 (1942).

THE SPREAD OF AN EPIDEMIC

BY EDWIN B. WILSON AND JANE WORCESTER

HARVARD SCHOOL OF PUBLIC HEALTH

Communicated September 11, 1945

The Frost-Soper theory of the rise and fall of an epidemic in a population[1] can at best represent only a part of the phenomenon of epidemics, for they spread from population to population as well as from individual infectious cases to susceptible individuals within a population. The fundamental hypothesis or law, namely,

$$C(t) = rS(t)C(t - \tau) \tag{1}$$

where C is the new-case rate, S the number of susceptibles, r a factor of proportionality and τ the time lag between infection and infectiousness may be very naturally extended to represent what hypothetically might happen between two populations S and S' of susceptibles by writing

$$C(t) = rS(t)C(t - \tau) + r''S(T)C'(t - \tau), \tag{2}$$

$$C'(t) = r'S'(t)C'(t - \tau) + r'''S'(t)C(t - \tau). \tag{2'}$$

It would indeed be only natural to argue that the new-case rate C among the susceptibles S should be the sum of what it would be due to the infectious rate within S and of a similar term due to the infectious rate in the other population of which a number of infectious individuals proportional to the total number within S' come in contact with the susceptibles S.

It may first be observed that if the two susceptible populations S and S' were merely ideal subdivisions of one and the same population S_0 so that $S = pS_0$, $S' = qS_0$, $p + q = 1$, and $C = pC_0$, $C' = qC_0$, the rate of intermixture of the infectious $C(t - \tau)$ with the susceptibles S or S' being the same as the rate of intermixture of the infectious $C(t - \tau)$ with S_0, and the

same for $C'(t - \tau)$, we should have $r = r' = r'' = r'''$ and the two equations would become

$$pC_0(t) = rp^2S_0(t)\ C_0(t - \tau) + rpqS_0(t)\ C_0(t - \tau),$$
$$qC_0(t) = rq^2S_0(t)\ C_0(t - \tau) + rpqS_0(t)\ C_0(t - \tau).$$

Each of these equations is equivalent to

$$C_0(t) = rS_0(t)C_0(t - \tau)$$

which is that of the whole epidemic within the whole population of susceptibles.

If, however, one should at any instant actually separate the population S_0 of susceptibles into two parts $S = pS_0$, $S' = qS_0$ and the infectious cases $C_0(t - \tau) = pC_0(t - \tau)$, $C'(t - \tau) = qC_0(t - \tau)$, imagining that the rate of intermixture of each set of infectious cases with its corresponding set of susceptibles has not been thereby changed but that the intermixture of the infectious $C(t - \tau)$ with susceptibles $S'(t)$ and of $C'(t - \tau)$ with $S(t)$ has been completely stopped by this separation one would have $r = r'$ but $r'' = r''' = 0$ and the two equations for the two separate epidemics must relapse to the form

$$C(t) = rS(t)C(t - \tau) \qquad \text{and } C'(t) = rS'(t)C'(t - \tau).$$

As the values of $S(t)$ and $C(t - \tau)$ are initially pS_0 and $pC_0(t - \tau)$, the value of the new-case rate $C(t)$ has been multiplied by the factor p^2 instead of by p and $C'(t)$ has been multiplied by q^2 and the total new-case rate $C(t) + C'(t)$ has initially been multiplied by $p^2 + q^2$ which at most is equal to $1/2$. The subsequent course of the two epidemics, each being regulated by its own equation, will not result in the total subsequent number of cases being the fraction $p^2 + q^2$ of the value of the total number which would have occurred if the separation had not been made, but may be something very different.[2] The principle, however, that one tends to break up an epidemic by separating the susceptibles into smaller groups even if the intimacy of intermixture of the infectious and the susceptibles within each group is not reduced is well known.

It is obvious that any argument which would lead to equations (2) and (2′) for the regulation of an *a priori* epidemic in two intercommunicating populations would immediately generalize into the formula

$$C_i(t) = S_i(t) \sum_{j=1}^{n} r_{ij}C_j(t - \tau), \qquad i = 1, 2, \ldots, n, \qquad (3)$$

for any one of n intercommunicating populations. If there are no new susceptibles coming into the population $C = -dS/dt$ and the integral

$$\log S_i(t) = \sum_{j=1}^{n} r_{ij}S_j(t - \tau) - K_i, \qquad i = 1, 2, \ldots, n, \quad (4)$$

may be obtained at once and, given the value of r_{ij} and of $S_i(t_0)$ and $S_i(t_0 - \tau)$ at the start, one can compute step by step all the n epidemics in the populations. It would seem that to consider intercommunicating populations under such an *a priori* law would be one step toward realism beyond the assumption of the simple mass law for a single population.

We shall take as an illustration two populations $S = 2000$, $S' = 1200$, $r = r' = 0.001$, $r'' = r''' = 0.0001$, so that the rate of intermixture of infectious from one population and susceptibles of the other is only one-tenth the rate of intermixture of the infectious of either population with that population. We start with the next value of S as 1998 and of S' as unchanged at 1200 to determine the constants. If we use $x = 0.001S$ and $x' = 0.001S'$ we have

$$\log x_{T+1} = x_T + 0.1\, x'_T - 1.42785,$$
$$\log x'_{T+1} = x'_T + 0.1\, x_T - 1.21768.$$

In successive generations the cases in the first and second populations and in the total are given in the first three columns of table 1.

From the above calculation made with the law of mass action but with allowance for a smaller rate of contact between the two populations than within each population, we may consider three theoretical epidemics, either that in the first or that in the second or that in the total population, and this is, in a simple idealized form, just what is actually done in the discussion of epidemics of infectious disease where the health officer may be interested in part of a population as well as in the total. If we consider, as Soper did, that the ratio of case rates at the center of successive incubation periods may be approximated by the ratio of the cases during those successive periods, we may obtain from the values of S (not given in table 1) and from the cases as given[3] values for m on the assumption that $p = 1$ (in our previous notation); these values of m are given in the last three columns of table 1 for the epidemics defined by the respective first three columns.

What is observed from the table is that the three epidemics are quite typical but that if any one of the three be used to determine the value of m, assuming $p = 1$, the value of m is not constant; for the epidemic in the first population m decreases throughout (the infectivity measured as the reciprocal of m increases), for that in the second population m increases throughout (the infectivity decreases), and for that in the total m increases to a maximum and then decreases (the infectivity decreases to a minimum and then increases). If, relying on the simple equation $C(t)/C(t - \tau) = S(t)/m$ one should take as the value of m the number of susceptibles when $C(t) = C(t - \tau)$, estimating the case rates as best he could, he would find something like $m_{\text{I}} = 985$, $m_{\text{II}} = 830$, $m_{\text{tot.}} = 1880$. If one should get the value of m/p from the ratio of total cases squared to eight times the peak

cases he would find something like these three respective values: 1050, 520, 1590. With the values of m found previously he would then estimate $p_{I} = 0.94$, $p_{II} = 1.6$, $p_{tot.} = 1.2$. Thus he would infer that $p = 1$ did not satisfy the relations. If these values of p were used to compute m the results would be different from those in the table.[4]

TABLE 1

A THEORETICAL EPIDEMIC IN EACH OF TWO INTERCOMMUNICATING POPULATIONS AND IN THE WHOLE POPULATION WITH THE VALUES OF SOPER'S m FOR EACH

CASES I	CASES II	TOTAL	m_{I}	m_{II}	$m_{tot.}$
2.1	0.0	2.0	...	...	...
4.0	0.2	4.2	1001	...	1523
8.0	0.8	8.7	993	...	1547
15.9	1.9	17.8	989	...	1566
31.5	4.1	35.6	988	543	1578
60.8	8.6	69.4	987	574	1593
112.3	17.3	129.6	986	593	1617
190.3	32.8	223.2	986	616	1660
277.2	57.3	334.6	986	648	1730
319.9	87.8	407.8	985	698	1824
274.0	111.7	385.7	974	765	1899
174.7	113.9	288.6	951	836	1900
89.9	94.0	183.9	918	893	1841
41.5	65.5	107.0	876	930	1776
18.7	40.6	59.3	835	950	1732
8.5	23.4	32.0	796	961	1705
4.0	13.0	17.0	787	969	1693
1.9	7.0	8.9	...	973	1686
1.0	3.7	4.7	...	973	1681
0.5	2.0	2.4	...	979	1678
0.2	1.0	1.3	...	...	..
0.1	0.6	0.7	...	...	...

From such a simple hypothetical example one may infer that even if epidemics in single homogeneous populations when isolated should be accountable by the law of mass action with $p = 1$, epidemics in groups of intercommunicating populations could not be so accounted for. As any real population is in fact not homogeneous but at best must be considered as formed of parts with different rates of intermixing of infectious and susceptibles both as within and as between parts, there is ample justification for generalizing the law of mass action to other values of p than 1 as an empirical law if thereby the observed relations are notably better reproduced.

We may observe that if populations are intercommunicating the whole question of periodicity becomes more complicated than in a single homo-

geneous population. If we take equations (2) and (2′) and replace C by $A - dS/dt$ where A is the rate of accession of susceptibles we have

$$A - \frac{dS}{dt} = S\left[r\left(A - \frac{dS}{dt}\bigg|_{t-\tau}\right) + r''\left(A' - \frac{dS'}{dt}\bigg|_{t-\tau}\right)\right],$$
$$A' - \frac{dS'}{dt} = S'\left[r'\left(A' - \frac{dS'}{dt}\bigg|_{t-\tau}\right) + r'''\left(A - \frac{dS}{dt}\bigg|_{t-\tau}\right)\right].$$

For the steady state to be possible in each population with $S = m$ and $S' = m'$ it is necessary that

$$A = m(rA + r''A') \qquad \text{and } A' = m'(r'A' + r'''A).$$

For an infinitesimal disturbance one has $S = m(1 + u)$, $S' = m'(1 + u')$. The symbolic form of the linear equations regulating the infinitesimal epidemics is therefore

$$[D - rmDe^{-\tau D} + rA + r''A']u - r''m'De^{-\tau D}u' = 0$$
$$-r'''mDe^{-\tau D}u + [D - r'm'De^{-\tau D} + r'A' + r'''A]u' = 0.$$

The periods (and dampings) will thus be had from the equation

$$[D - rmDe^{-\tau D} + rA + r''A'][D - r'm'De^{-\tau D} + r'A' + r'''A] - r''r'''mm'D^2e^{-2\tau D} = 0.$$

The simplest approximation to make would be to assume $e^{-\tau D} = 1 - \tau D$. This would lead to a biquadratic equation for D and thus normally to a pair of periods with different periodic times. Presumably the damping would be neglected. It is well known that the superposition of a pair of periodic terms may make a very irregular resultant variation.

Now, the variation of measles over the different wards of a city or over the different suburbs of a large metropolitan area is such, on the record, as to show both some degree of independence of the disease in the different sections and some degree of interdependence between sections.[5] While the general theory, whether of rise or fall or of periodicity, built upon some simple empirical law may throw much light on the observed phenomena, no such theory can be expected to fit them exactly. It seems to be indicated that one must study in detail the laws of spread, whether in the field or in the laboratory, before he can be finally satisfied with the theoretical discussion of epidemics.

[1] Wilson, E. B., and Burke, M. H., these PROCEEDINGS, 28, 361–367 (1942); 29, 43–48 (1943); and Wilson, E. B., and Worcester, J., *Ibid.*, 30, 37–44 and 264–269 (1944); 31, 24–34, 109–116, 142–147, 203–208, 294–298 (1945).

[2] If we take an epidemic starting with 2000 susceptibles and $r = 0.001$ so that the next number of susceptibles is 1998, it being assumed that there are no new susceptibles coming into the population, the rule for successive generations is $\log x_{T+1} = x_T - 1.30785$

where $x_T = 0.001S_T$. The epidemic then goes on with these successive values of S, the cases being the differences between the values of S_0,

$S = 2000.00$,		1998.00,		1994.01,		1986.08,		1970.40,
Cases	2.00,		3.99,		7.93,		15.68.	

If the calculation be carried on indefinitely we should have a final asymptotic $S = 406$ and total cases of 1594. If, however, the population be cut in two at the point reached and we consider $S_T = 985.20$ susceptibles with 7.84 infectious cases and $r = 0.001$, the number of new cases would be 7.72 with $S_{T+1} = 977.48$ so that $\log x_{T-1} = x_T - 1.00797$ would be the equation to use in the stepwise calculation. The argument could also be given in this form: The new number of cases if the separation had not been made would have been 30.66; with the separation it should have been cut in half to 15.33 and for each half it would have been 7.66; then taking $S_T = 985.20$, $S_{T+1} = 977.54$ we should determine $\log x_{T+1} = x_T - 1.00792$ for the stepwise calculation. Probably the best value of x_{T+1} would be between 0.97748 and 0.97754 and the best value of the constant between 1.00797 and 1.00792. To get the best new equation for each half of the epidemic, particularly in cases where the separation is made when the ratio of new cases to susceptibles is higher than in this example and where the two modes of calculation would differ more seriously, would require very careful analysis involving expansion into series in the time so as to distinguish adequately between cases and case rates. If the calculation of the example just taken be made with the second equation obtained the total number of cases in each half after the time of separation is 105 instead of 1565 for the whole after the corresponding time; it will be noted that $2 \times 105 = 210$ is decidedly less than $p^2 + q^2 = 1/2$ of 1565. In fact the separation has stopped the epidemic so completely that the number of cases decreases from the time of separation.

[3] There are a variety of formulae which may be developed for getting the ratio of two ordinates in terms of successive areas. The one used by Soper is perhaps the simplest. As, however, the epidemic case-rate curve for a single population much resembles the sech^2 curve we might consider that for the ratio $C(t - \tau/2)/C(t + \tau/2)$ one could fit a sech^2 curve to C from $t - 3\tau/2$ to $t + 3\tau/2$ using K_-, K_0 and K_+ as the cases in the intervals $t - 3\tau/2$ to $t - \tau/2$, $t - \tau/2$ to $t + \tau/2$, and $t + \tau/2$ to $t + 3\tau/2$. The simple resulting formula is

$$\frac{C(t - \tau/2)}{C(t + \tau/2)} = \frac{m(t + \tau/2)}{S(t + \tau/2} = \frac{K_-}{K_+}\frac{K_0 + K_+}{K_0 + K_-}$$

and it has been used for the actual calculation of m instead of Soper's. The difference between the two is, however, tolerably small.

[4] We might, as in a previous article, keep p constant and compute the variation of m throughout the epidemic, or we might keep m constant and compute the variation of p, or finally we might let both m and p be determined from successive pairs of data. In using data from actual epidemics we found a great deal of scattering in the successive values of m and p but apparently a tendency for m to decrease throughout the epidemic and for p first to decrease from large values and then to increase to large values. If we make the corresponding calculations on these theoretical epidemics we find that m remains practically constant throughout the major course of the epidemic (at the ends enough places were lost to make the figures unreliable) and the value of p decreases throughout the epidemic whether this be that in either part of the population or that in the whole. We have assumed there are no recruits to the susceptibles, for without this assumption we cannot get exact integrals; clearly when there are recruits, as there probably are in real epidemics, they may well make a considerable difference with calcu-

lations when the number of new cases is less or only slightly greater than the number of new susceptibles during each incubation period.

[6] Wilson, E. B., Bennett, C., Allen, M., and Worcester, J., *Proc. Amer. Phil. Soc.*, **80**, 357–476 (1939). See particularly Appendix II, Localness of Measles, 469–476.

A TOPOLOGY FOR THE SET OF PRIMITIVE IDEALS IN AN ARBITRARY RING

BY N. JACOBSON

DEPARTMENT OF MATHEMATICS, JOHNS HOPKINS UNIVERSITY

Communicated July 11, 1945

1. In a recent paper we have called a ring $\mathfrak{A}$ primitive if $\mathfrak{A}$ contains a maximal right ideal $\mathfrak{J}$ whose quotient $\mathfrak{J}:\mathfrak{A} = 0$.[1] In general if $\mathfrak{J}$ is any right ideal $\mathfrak{J}:\mathfrak{A}$ is the totality of elements b such that $xb \epsilon \mathfrak{J}$ for all x in $\mathfrak{A}$. $\mathfrak{J}:\mathfrak{A}$ is a two-sided deal and if $\mathfrak{A}$ has an identity, $\mathfrak{J}:\mathfrak{A}$ is the largest two-sided ideal of $\mathfrak{A}$ contained in $\mathfrak{J}$. The primitive rings appear to play the same rôle in the general structure theory of rings that is played by simple rings in the classical theory of rings that satisfy the descending chain condition for one-sided ideals. Corresponding to the Wedderburn-Artin structure theorem on simple rings satisfying the descending chain condition we have the theorem that if $\mathfrak{A}$ is a primitive ring $\neq 0$, $\mathfrak{A}$ is isomorphic to a dense ring of linear transformations in a suitable vector space over a division ring.

We shall call a two-sided ideal $\mathfrak{B}$ in $\mathfrak{A}$ a *primitive ideal* if $\mathfrak{B} \neq \mathfrak{A}$ and $\mathfrak{A} - \mathfrak{B}$ is a primitive ring. It is known that if $\mathfrak{A}$ is not a radical ring then $\mathfrak{A}$ contains primitive ideals. Moreover, in this case the intersection $\Pi\mathfrak{B}$ of all the primitive ideals in $\mathfrak{A}$ coincides with the radical $\mathfrak{R}$ of $\mathfrak{A}$. In particular if $\mathfrak{A}$ is semi-simple, $\Pi\mathfrak{B} = 0$. If $\mathfrak{A}$ is a ring with an identity, $\mathfrak{A}$ is not a radical ring. Hence if $\mathfrak{C}$ is any two-sided ideal $\neq \mathfrak{A}$ in $\mathfrak{A}$, $\mathfrak{A} - \mathfrak{C}$ contains a primitive ideal $\mathfrak{B} - \mathfrak{C}$. Since $\mathfrak{A} - \mathfrak{B} \cong (\mathfrak{A} - \mathfrak{C}) - (\mathfrak{B} - \mathfrak{C})$, $\mathfrak{B}$ is primitive in $\mathfrak{A}$. Thus if $\mathfrak{A}$ is a ring with an identity any two-sided ideal $\mathfrak{C} \neq \mathfrak{A}$ can be imbedded in a primitive ideal.

Any commutative primitive ring is a field. Consequently any primitive ideal in a commutative ring is maximal. On the other hand, in a non-commutative ring there may exist primitive ideals that are not maximal. For example, the ring $\mathfrak{L}$ of all linear transformations in an infinite dimensional vector space $\mathfrak{R}$ over a division ring is primitive. Hence (0) is a primitive ideal. However, $\mathfrak{L}$ contains as a proper two-sided ideal the set $\mathfrak{F}$ of finite valued linear transformations in $\mathfrak{R}$.

A simple ring is either primitive or a radical ring. In particular any simple ring with an identity is primitive. These remarks imply that a maximal two-sided ideal $\mathfrak{B}$ such that $\mathfrak{A} - \mathfrak{B}$ is not a radical ring is primitive. If $\mathfrak{A}$ is a ring with an identity, any maximal two-sided ideal in $\mathfrak{A}$ is primitive.

In this note we shall define a topology for the set of primitive ideals of any ring. The space determined in this way appears to be an important invariant of the ring. We hope to discuss its rôle in the general structure theory in greater detail at a later date.

2. Let S be the set of primitive ideals in the ring $\mathfrak{A}$. S is vacuous if and only if $\mathfrak{A}$ is a radical ring. If A is a non-vacuous subset of S we let $\mathfrak{D}_A$ denote the intersection of all the primitive ideals $\mathfrak{B} \epsilon A$. We now define the closure $\bar{A}$ of A to be the totality of primitive ideals $\mathfrak{C}$ such that $\mathfrak{C} \geqq \mathfrak{D}_A$. It is clear that

1. $\bar{A} \geqq A$.
2. $\bar{\bar{A}} = \bar{A}$.

We wish to prove next that

3. $\overline{A \vee B} = \bar{A} \vee \bar{B}$.[2]

Proof. Let $\mathfrak{B} \epsilon \bar{A} \vee \bar{B}$, say $\mathfrak{B} \epsilon \bar{A}$. Then $\mathfrak{B} \geqq \mathfrak{D}_A$. Hence $\mathfrak{B} \geqq \mathfrak{D}_A \wedge \mathfrak{D}_B = \mathfrak{D}_C$ where $C = A \vee B$. Thus $\mathfrak{B} \epsilon \overline{A \vee B}$. Suppose next that $\mathfrak{B} \bar{\epsilon} \bar{A} \vee \bar{B}$. Then $\mathfrak{B} \not\geqq \mathfrak{D}_A$ and $\mathfrak{B} \not\geqq \mathfrak{D}_B$. We consider now the primitive ring $\mathfrak{A} - \mathfrak{B}$. The two-sided ideals $(\mathfrak{D}_A + \mathfrak{B}) - \mathfrak{B}$ and $(\mathfrak{D}_B + \mathfrak{B}) - \mathfrak{B}$ are $\neq 0$ in $\mathfrak{A} - \mathfrak{B}$. Hence the product $[(\mathfrak{D}_A + \mathfrak{B}) - \mathfrak{B}][(\mathfrak{D}_B + \mathfrak{B}) - \mathfrak{B}] \neq 0$.[3] It follows that $\mathfrak{D}_A\mathfrak{D}_B \not\leqq \mathfrak{B}$. Hence also $\mathfrak{D}_A \wedge \mathfrak{D}_B \not\leqq \mathfrak{B}$ and so $\mathfrak{B} \bar{\epsilon} \overline{A \vee B}$.

For the vacuous set ω we define

4. $\bar{\omega} = \omega$.

The properties 1–4 show that S *is a topological space relative to the closure operation* $A \to \bar{A}$. In the special case of a Boolean ring this topology is due to Stone.[4] It has also been introduced by Gelfand and Silov in commutative normed rings.[5] We shall call the topological space S the *structure space* of the ring $\mathfrak{A}$.

If $\mathfrak{B}$ is a primitive ideal in $\mathfrak{A}$, $\mathfrak{B}$ is a point in S. The closure of the set $\{\mathfrak{B}\}$ is the totality of primitive ideals $\mathfrak{C}$ such that $\mathfrak{C} \geqq \mathfrak{B}$. Hence if $\overline{\{\mathfrak{B}_1\}} = \overline{\{\mathfrak{B}_2\}}$ then $\mathfrak{B}_1 = \mathfrak{B}_2$. This shows that S is a T_0-space. In general S is not a T_1-space. For we have seen that there exist primitive rings $\mathfrak{A}$ that are not simple. If $\mathfrak{A}$ is a ring with an identity of this type and $\mathfrak{B}$ is a proper two-sided ideal in $\mathfrak{A}$, $\mathfrak{B}$ can be imbedded in a primitive ideal $\mathfrak{C}$. Then $\overline{\{(0)\}}$ contains $\mathfrak{C} \neq 0$. The subspace M of S of primitive ideals that are maximal is clearly a T_1-space. If $\mathfrak{A}$ is commutative $S = M$ is a T_1-space. However, even in this case S need not be a T_2- (or Hausdorff) space. An example of a normed ring of this type has been given by Gelfand and Silov.[5] A simpler one is the following:

Example. Let $\mathfrak{A} = J$ the ring of integers. The primitive ideals are the prime ideals (p). Since the intersection of an infinite number of prime ideals is the 0-ideal, $\bar{A} = S$ for any infinite set A. If A is finite, $\bar{A} = A$. Hence the open sets $\neq \omega$, $\neq S$ are the complements of finite sets. Any two open sets $\neq \omega$ have a non-vacuous intersection and so the Hausdorff separation property does not hold.

3. Let $\mathfrak{A}_1$ be an arbitrary two-sided ideal in $\mathfrak{A}$ and let S_1 denote the closed set in S consisting of the primitive ideals $\mathfrak{B}$ of $\mathfrak{A}$ that contain $\mathfrak{A}_1$. If $\mathfrak{B} \epsilon S_1$ then $\mathfrak{B} - \mathfrak{A}_1$ is a primitive ideal in $\mathfrak{A} - \mathfrak{A}_1$ and any primitive ideal in $\mathfrak{A} - \mathfrak{A}_1$ is obtained in this way. The correspondence $\mathfrak{B} \rightarrow \mathfrak{B} - \mathfrak{A}_1$ is $(1 - 1)$ between the subspace S_1 of S and the structure space T of the ring $\mathfrak{A} - \mathfrak{A}_1$. Since this correspondence preserves intersection it is a homeomorphism between S_1 and T.

Suppose in particular that $\mathfrak{A}_1 = \mathfrak{R}$ the radical of $\mathfrak{A}$. Then every primitive ideal of $\mathfrak{A}$ contains $\mathfrak{R}$. Hence $S_1 = S$, and we see that *the structure space of* $\mathfrak{A}$ *is homeomorphic to the structure space of the semi-simple ring* $\mathfrak{A} - \mathfrak{R}$.

We return to the general case in which $\mathfrak{A}_1$ is arbitrary and we now consider the open set S_1' of primitive ideals that do not contain $\mathfrak{A}_1$. Let $\mathfrak{C} \epsilon S_1'$. Then $\mathfrak{C} \wedge \mathfrak{A}_1$ is a two-sided ideal $\neq \mathfrak{A}_1$ in $\mathfrak{A}_1$ and $\mathfrak{A}_1 - (\mathfrak{C} \wedge \mathfrak{A}_1) \cong (\mathfrak{C} + \mathfrak{A}_1) - \mathfrak{C}$. The latter ring is a two-sided ideal in the primitive ring $\mathfrak{A} - \mathfrak{C}$. Hence it is primitive.[6] Thus $\mathfrak{C} \wedge \mathfrak{A}_1$ is in the structure space U of the ring $\mathfrak{A}_1$. Let A be a subset of S_1' and let $\mathfrak{C} \epsilon \bar{A} \wedge S_1'$ so that $\mathfrak{C}$ is in the closure of A in the subspace S_1'. If $\mathfrak{D}_A$ is the intersection of the primitive ideals in A then $\mathfrak{C} \geqq \mathfrak{D}_A$ but $\mathfrak{C} \not\geqq \mathfrak{A}_1$. Let B be the subset of U of ideals $\mathfrak{B} \wedge \mathfrak{A}_1$ where $\mathfrak{B} \epsilon A$. The intersection of all of these ideals is the ideal $\mathfrak{D}_A \wedge \mathfrak{A}_1$. Since $(\mathfrak{C} \wedge \mathfrak{A}_1) \geqq (\mathfrak{D}_A \wedge \mathfrak{A}_1)$, $\mathfrak{C} \wedge \mathfrak{A}_1$ is in $\bar{B}$. Hence the mapping that we have defined between S_1' and the subspace of U is a continuous one.

4. Let $\{F_\alpha\}$ be a set of closed sets in the space S. As before let $\mathfrak{D}_{F_\alpha}$ denote the two-sided ideal of elements common to all the $\mathfrak{B} \epsilon F_\alpha$. Suppose that the intersection $\Pi F_\alpha \neq \omega$ and let $\mathfrak{B}$ be a point in this intersection. Then $\mathfrak{B} \geqq \mathfrak{D}_{F_\alpha}$ for all α. Hence $\mathfrak{B}$ contains the two-sided ideal $\Sigma\mathfrak{D}_{F_\alpha}$ generated by the $\mathfrak{D}_{F_\alpha}$. The converse follows by retracing the steps of this argument. We therefore have the

LEMMA 1. *If* $\{F_\alpha\}$ *is a set of closed sets in* S, $\Pi F_\alpha \neq \omega$ *if and only if* $\Sigma\mathfrak{D}_{F_\alpha}$ *can be imbedded in a primitive ideal.*

If $\mathfrak{A}$ is a ring with an identity any two-sided ideal $\neq \mathfrak{A}$ can be imbedded in a primitive ideal. Hence we have the

Corollary. If $\mathfrak{A}$ *is a ring with an identity and* $\{F_\alpha\}$ *is a set of closed sets in* S *then* $\Pi F_\alpha \neq \omega$ *if and only if* $\Sigma\mathfrak{D}_{F_\alpha} \neq \mathfrak{A}$.

If $\mathfrak{A}$ has an identity and $\Sigma\mathfrak{D}_{F_\alpha} = \mathfrak{A}$, $1 = d_1 + \ldots + d_r$ where $d_i \epsilon \mathfrak{D}_{F_i}$. Hence also $\Sigma\mathfrak{D}_{F_i} = \mathfrak{A}$. The corollary therefore shows that if $\Pi F_\alpha = \omega$

then there is a finite set of closed sets F_i such that $\Pi F_i = \omega$. We therefore have the following

THEOREM 1. *The structure space of a ring with an identity is bicompact.*

Suppose now that $\mathfrak{A}$ is a semi-simple ring. Let $\mathfrak{A}$ be decomposable as a direct sum $\mathfrak{A}_1 \oplus \mathfrak{A}_2$ of the two-sided ideals $\mathfrak{A}_i \neq 0$. Let S_i be the closed subset of S consisting of the primitive ideals $\mathfrak{B}$ containing the ideal $\mathfrak{A}_i$. Since $\mathfrak{A} - \mathfrak{A}_1 \cong \mathfrak{A}_2$ a semi-simple ring, $S_1 \neq \omega$. Similarly $S_2 \neq \omega$. Also the semi-simplicity of $\mathfrak{A} - \mathfrak{A}_i$ implies that $\mathfrak{A}_i$ is the intersection of all the $\mathfrak{B}$ in S_i. Since $\mathfrak{A}_1 + \mathfrak{A}_2 = \mathfrak{A}$ the lemma implies that $S_1 \wedge S_2 = \omega$. Now let $\mathfrak{B}$ be any primitive ideal in $\mathfrak{A}$. We assert that either $\mathfrak{B} \geqq \mathfrak{A}_1$ or $\mathfrak{B} \geqq \mathfrak{A}_2$. For otherwise $\mathfrak{A}_i + \mathfrak{B} > \mathfrak{B}$. The ideals $(\mathfrak{A}_i + \mathfrak{B}) - \mathfrak{B}$ are $\neq 0$ in the primitive ring $\mathfrak{A} - \mathfrak{B}$. Since $[(\mathfrak{A}_1 + \mathfrak{B}) - \mathfrak{B}][(\mathfrak{A}_2 + \mathfrak{B}) - \mathfrak{B}] = 0$ this is impossible. We have therefore proved our assertion. Evidently it is equivalent to the relation $S_1 \vee S_2 = S$. Thus S is disconnected into the two components S_1 and S_2.

Conversely suppose that $S = S_1 \vee S_2$ where the S_i are closed sets $\neq \omega$ such that $S_1 \wedge S_2 = \omega$. We assume also that $\mathfrak{A}$ has an identity. Let $\mathfrak{A}_i = \mathfrak{D}_{S_i}$. Then $\mathfrak{A}_1 + \mathfrak{A}_2 = \mathfrak{A}$. Since $\mathfrak{A}$ is semi-simple $\mathfrak{A}_1 \wedge \mathfrak{A}_2 = \mathfrak{D}_S = 0$. If $\mathfrak{A}_i = 0$, $S = \bar{S}_i = S_i$ contrary to $S_1 \neq \omega$ and $S_2 \neq \omega$. A part of our result is the following

THEOREM 2. *If $\mathfrak{A}$ is a semi-simple ring with an identity, $\mathfrak{A} = \mathfrak{A}_1 \oplus \mathfrak{A}_2$ where the $\mathfrak{A}_i$ are two-sided ideals $\neq 0$ if and only if S is not connected.*

5. If S' is a completely regular bicompact space and $\mathfrak{S}$ is the ring of real-valued (complex valued) continuous functions on S' then it has been shown by Gelfand and Silov that the structure space S of $\mathfrak{S}$ is homeomorphic to S'.[7] If S' is a bicompact totally disconnected space and $\mathfrak{S}$ is the ring of continuous functions on S' having values in the field of residues mod 2, then by a result of Stone's the structure space of $\mathfrak{S}$ is homeomorphic to S'.[8] We conclude this note by giving another example of this type based on an arbitrary totally disconnected bicompact space S' and an arbitrary division ring $\mathfrak{K}'$.

We consider any decomposition of S' into a finite number of components (non-overlapping open and closed sets) S'_i and we choose corresponding elements $k_i \epsilon \mathfrak{K}'$. We define a function $f(x)$ by setting $f(x_i) = k_i$ for x_i in S'_i. A function of this type will be called a *finite decomposition function.* The totality $\mathfrak{S}$ of these functions is a ring under the ordinary operations of addition and multiplication. $\mathfrak{S}$ contains the subring $\mathfrak{K}$ of constant functions, isomorphic to $\mathfrak{K}'$ and $\mathfrak{S}$ is commutative if and only if $\mathfrak{K}'$ is commutative. The constant 1 acts as an identity in $\mathfrak{S}$. Since S' is totally disconnected, for any two points $a \neq b$ in S' there is an $f(x) \epsilon \mathfrak{S}$ such that $f(a) \neq f(b)$. Let $\mathfrak{A}$ be any subring of $\mathfrak{S}$ having this property and containing $\mathfrak{K}$. We shall sketch a proof of the fact that the space M of maximal two-sided ideals of $\mathfrak{A}$ is homeomorphic to S'.

LEMMA 2. *If F' is a closed subset of S' and a is a point $\bar{\epsilon}$ F' then there exists a function $\varphi(x) \epsilon \mathfrak{A}$ such that $\varphi(y) = 0$ for all $y \epsilon F'$ but $(\varphi a) \neq 0$.*

Our assumptions imply that for each $y \epsilon F'$ there is an $f_y \epsilon \mathfrak{A}$ such that $f_y(y) = 0$ but $f_y(a) \neq 0$. The set $Z'(f_y)$ of zeros of f_y is open and the totality of these sets covers F'. Let $Z'(f_{y_1}), \ldots, Z'(f_{y_m})$ be a finite subset of these sets covering F'. Then $\varphi(x) = f_{y_1}(x) \ldots f_{y_m}(x)$ has the required properties.

If $a \epsilon S'$ we let $\mathfrak{B}_a$ denote the totality of functions $g(x) \epsilon \mathfrak{A}$ such that $g(a) = 0$. It is easy to see that $\mathfrak{B}_a$ is a maximal two-sided ideal in $\mathfrak{A}$ and that $\mathfrak{A} - \mathfrak{B}_a \cong \mathfrak{K}$. Our conditions imply that if $a \neq b$ then $\mathfrak{B}_a \neq \mathfrak{B}_b$.

LEMMA 3. *If $\mathfrak{B}$ is a two-sided ideal $\neq \mathfrak{A}$, then there exists a point a such that $g(a) = 0$ for all $g \epsilon \mathfrak{B}$.*

Let $Z'(f)$ be the set of zeros of f. $Z'(f)$ is closed. If the lemma is false, the intersection $\Pi Z'(f)$ for all $f \epsilon \mathfrak{B}$ is vacuous. Hence there is a finite, number of functions $f_1, \ldots, f_r$ in $\mathfrak{B}$ such that $\Pi Z'(f_i) = \omega$. Hence if a is any point of S' at least one of the functions f_i does not vanish at a. Let $k_{1i}, \ldots, k_{n_i i}$ be the non-zero values taken on by f_i and form the function $\psi_i(x) = (f_i(x) - k_{1i})(f_i(x) - k_{2i}) \ldots (f_i(x) - k_{n_i i})$ and $\psi(x) = \psi_1(x)\psi_2(x) \ldots \psi_r(x)$. Then $\psi(x) \equiv 0$. The form of $\psi(x)$ shows that $0 = \psi(x) = k + g(x)$ where $k \neq 0$ and $g(x) \epsilon \mathfrak{B}$. Hence $\mathfrak{B}$ contains a constant function $\neq 0$ and $\mathfrak{B} = \mathfrak{A}$ contrary to asumption.

This lemma shows that $\mathfrak{B} \leqq \mathfrak{B}_a$ for some a. If $\mathfrak{B}$ is maximal $\mathfrak{B} = \mathfrak{B}_a$. Thus the correspondence $a \rightarrow \mathfrak{B}_a$ is $(1 - 1)$ between S' and the space M of maximal two-sided ideals in $\mathfrak{A}$.

Let F' be a closed subset of S' and let F be the corresponding set in M. The intersection $\mathfrak{D}_F = \Pi\mathfrak{B}_a$ for all $a \epsilon F'$ is the set of functions f such that $f(a) = 0$ for all $a \epsilon F$. Let $\mathfrak{B}_b$ be a maximal two-sided ideal containing $\mathfrak{D}_F$. If $b \bar{\epsilon} F'$ there is a function ϕ such that $\phi(a) = 0$ for all a in F' but $\phi(b) \neq 0$. Then $\phi \epsilon \mathfrak{D}_F$ but $\bar{\epsilon}\, \mathfrak{B}_b$ contrary to $\mathfrak{B}_b \geqq \mathfrak{D}_F$. Hence $b \epsilon F'$ and $\mathfrak{B}_b \epsilon F$. Thus F is closed. Conversely let F be any closed set in M and let F' be the corresponding set in S'. It is easy to see that F' is the set of points y such that $f(y) = 0$ for all f in $\mathfrak{D}_F$. Thus F' is the intersection of closed sets and is therefore closed. Hence the correspondence $a \rightarrow \mathfrak{B}_a$ is a homeomorphism.

THEOREM 3. *Let S' be a totally disconnected bicompact space, $\mathfrak{K}'$ a division ring and $\mathfrak{A}$ a ring of $\mathfrak{K}'$-valued decomposition functions such that (1) $\mathfrak{A}$ contains the constants and (2) if $a \neq b$ in S', then $\mathfrak{A}$ contains a function f such that $f(a) \neq f(b)$. Then the space M of maximal two-sided ideals of $\mathfrak{A}$ is homeomorphic to S'.*

If $\mathfrak{K}'$ is commutative, $M = S$. It is an open question whether or not this is true in general. However, it is clear that $\bar{M} = S$ since the intersection of the set of maximal two-sided ideals is (0).

Theorem 3 shows that two rings $\mathfrak{A}_1$ and $\mathfrak{A}_2$ of the type considered are not isomorphic unless the spaces S_1' and S_2' are homeomorphic. Since the division ring $\mathfrak{K}_i'$ is isomorphic to $\mathfrak{A}_i - \mathfrak{B}_i$ for any maximal two-sided ideal

$\mathfrak{B}_i$ in $\mathfrak{A}_i$, the isomorphism of $\mathfrak{A}_1$ and $\mathfrak{A}_2$ implies that of $\mathfrak{K}_1'$ and $\mathfrak{K}_2'$. If $\mathfrak{A}_i = \mathfrak{S}_i$ the complete ring of finite decomposition functions then the converse holds. Hence necessary and sufficient conditions that $\mathfrak{S}_1$ and $\mathfrak{S}_2$ be isomorphic are that S_1' and S_2' be homeomorphic and that $\mathfrak{K}_1'$ and $\mathfrak{K}_2'$ be isomorphic.

[1] "The Radical and Semi-simplicity for Arbitrary Rings," *Amer. Jour. Math.*, **67**, 300–320 (1945). The results that we state without proof in this section can be found in the above-mentioned paper.

[2] We use the notations $\vee$ and $\wedge$, respectively, for the logical sum and logical product of a finite number of sets.

[3] Loc. cit. in reference 1, p. 316.

[4] "Applications of the Theory of Boolean Rings to General Topology," *Trans. Amer. Math. Soc.*, **41**, 375–481 (1937).

[5] "Über verschiedene Methoden der Einführing der Topologie in die Menge der maximalen ideale eines normierten Ringes," *Math. Sbornik*, **51**, 25–39 (1941).

[6] Loc. cit. in reference 1, p. 313.

[7] Loc. cit. in reference 5, pp. 27–30.

[8] Loc. cit. in reference 4, p. 380. It is assumed here that the field of residues mod 2 is endowed with a T_1-topology. This means that each point is both open and closed.

CONVERSE OF PTOLEMY'S THEOREM ON STEREOGRAPHIC PROJECTION

By Edward Kasner and John De Cicco

Departments of Mathematics, Columbia University and Illinois Institute of Technology

Communicated August 22, 1945

1. *The Problem.*—The famous theorem of Ptolemy on stereographic projection states that the perspectivity of a sphere from a point of the sphere upon a plane perpendicular to the diameter of the sphere determined by the given point is conformal. The question naturally arises as to the existence of any other surfaces for which there exists a perspectivity upon a plane which is conformal. We prove that there does not exist any other surface with this property (except for the obvious case of a parallel plane). Our result may be stated as follows:

Theorem. *If a perspectivity of a surface from a fixed point to a fixed plane is conformal, then the surface must be a sphere; furthermore, the sphere must pass through the fixed point and have its center on the perpendicular from the fixed point to the fixed plane.*

Thus the only perspective conformalities upon a plane are Ptolemy's stereographic projection of the sphere (or the limiting case of the plane).

2. *Beginning of the Proof of the Theorem.*—Let (x, y, z) denote cartesian

coördinates of a point and let $p = \partial z/\partial x$ and $q = \partial z/\partial y$. Then $(p, q, -1)$ are the direction numbers of any normal to a surface.

We shall take the origin as the center of perspectivity and $z = c \neq 0$ as the given plane. Thus (X, Y, c) are the cartesian coördinates of any point in the given plane. Let the equation of the unknown surface be $z = f(x, y)$.

The perspectivity from the given point $(0, 0, 0)$ of the surface $(x, y, z = f(x, y))$ upon the given plane (X, Y, c) is given by

$$X = \frac{cx}{z}, \qquad Y = \frac{cy}{z}. \tag{1}$$

The differentials of this point transformation are

$$dX = \frac{c}{z^2}[(z - xp)dx - xqdy], \qquad dY = \frac{c}{z^2}[-ypdx + (z - yq)dy]. \tag{2}$$

The jacobian J of our transformation is

$$J = \frac{c^2}{z^4}[(z - xp)(z - yq) - xypq] = \frac{c^2}{z^3}(z - xp - yq). \tag{3}$$

The vanishing of the jacobian shows that $z = f(x, y)$ is a homogeneous function of first degree, according to Euler's theorem on homogeneous functions. This means that the surface cannot be a cone with vertex at the origin. Thus since we require regions to correspond, *the plane $z = 0$ and any cone with vertex at origin must be omitted from consideration.*

By equation (2), it is found that the linear element dS in the plane $z = c$, is given by

$$dS^2 = dX^2 + dY^2 = \frac{c^2}{z^4}\begin{bmatrix} \{z^2 - 2xzp + (x^2 + y^2)p^2\}dx^2 + \\ 2\{-z(yp + xq) + (x^2 + y^2)pq\}\,dxdy \\ + \{z^2 - 2yzq + (x^2 + y^2)q^2\}dy^2 \end{bmatrix}. \tag{4}$$

On the other hand, the linear element of the surface $z = f(x, y)$ is given by

$$ds^2 = (1 + p^2)dx^2 + 2pqdxdy + (1 + q^2)dy^2. \tag{5}$$

Since we require that the perspectivity of the surface $z = f(x, y)$ upon the plane $z = c$, be conformal, the scale $\sigma = dS/ds$ is independent of direction. Hence from equations (4) and (5), we must have

$$\frac{z^2 - 2xzp + (x^2 + y^2)p^2}{1 + p^2} = \frac{-z(yp + xq) + (x^2 + y^2)pq}{pq} = \frac{z^2 - 2yzq + (x^2 + y^2)q^2}{1 + q^2}. \tag{6}$$

Subtracting z^2 from each ratio of this proportion, we find

$$\frac{p[-2xz + (x^2 + y^2 - z^2)p]}{1 + p^2} = \frac{-z(yp + xq) + (x^2 + y^2 - z^2)pq}{pq} = \frac{q[-2yz + (x^2 + y^2 - z^2)q]}{1 + q^2}. \quad (7)$$

Let M and N denote the quantities (linear in p and q)

$$M = -2xz + (x^2 + y^2 - z^2)p, \quad N = -2yz + (x^2 + y^2 - z^2)q. \quad (8)$$

The equation (7) may be written as

$$\frac{pM}{1 + p^2} = \frac{qM + pN}{2pq} = \frac{qN}{1 + q^2}. \quad (9)$$

3. *The Special Cases, Real and Imaginary.—If p or q is zero, then both are zero, and we have a parallel plane as a solution.* For if $p = 0$, then $qM = 0$. If $q \neq 0$, then $M = -2xz = 0$ which is impossible. Hence from $p = 0$, follows the fact that $q = 0$ which proves our statement. The argument when $q = 0$ is similar. Henceforth we may assume that neither p nor q is zero.

If the quantity $p = \pm i$, then the quantity $q = \pm 1$ (or if $p = \pm 1$, then $q = \pm i$) and the solution of equation (9) is an imaginary plane through the origin which is excluded from consideration since the perspectivity (1) is degenerate in this case. Let us suppose that $p = i$. From equation (9), we find $-2qM = 0$. Since $q \neq 0$, it follows that $M = -2xz + i(x^2 + y^2 - z^2) = 0$. This yields the imaginary planes $z = ix \pm y$, which are solutions of equation (9). However, this is excluded since the perspectivity (1) is degenerate in this case. A similar argument will prove our statement for $p = -i$ or $q = \pm i$.

Let us suppose $p = 1$. From equation (9), we find immediately that $N = 0$ and hence either $M = -2xz + x^2 + y^2 - z^2 = 0$ or $1 + q^2 = 0$. Now the condition $M = -2xz + x^2 + y^2 - z^2 = 0$ yields a contradiction; for the partial derivative q from this is $q = y/(x + z)$ and from $N = 0$, we find it to be $q = y/x$. The equality of these expressions yields the contradiction $yz = 0$. Hence when $p = 1$, we must have $1 + q^2 = 0$. This gives the imaginary planes $z = x \pm iy$, which are omitted since the perspectivity (1) is degenerate. The argument is similar when $p = -1$ or $q = \pm 1$. This completes the proof of the preceding italicized statement. Henceforth we may assume that neither p nor q assumes any of the five values $0, \pm i, \pm 1$.

We shall consider as our final imaginary case, the condition $p^2 + q^2 = 0$. This leads to the imaginary planes $z = a(x \pm iy)$ through the origin, which are omitted from consideration since these make the perspectivity (1) degenerate. A solution of this partial equation of first order is $z = f(u)$ where $u = x + iy$

(the case where $z = f(v)$ where $v = x - iy$, is similar). Hence $z = f(u)$, $p = f_u$, $q = ip = if_u$. The equations (9) then become

$$\frac{M}{1 + p^2} = \frac{iM + N}{2ip^2} = \frac{iN}{1 - p^2} \tag{10}$$

because $p \neq 0$. Since $p \neq \pm 1, \pm i$, these equations are equivalent to the single condition

$$i(p^2 - 1)M - (p^2 + 1)N = 0. \tag{11}$$

Substituting equation (8) into this and simplifying, we obtain

$$(uv - z^2)p + z(up^2 - v) = 0. \tag{12}$$

Since z and p are both functions of u only, this identity in v leads to the single condition $uf_u(u) - f(u) = 0$. This yields the solution $z = au = a(x + iy)$, which is an imaginary plane through the origin. As stated above, this must be omitted from consideration as it makes the perspectivity (1) degenerate. Henceforth p and q cannot be any of the five numbers 0, $\pm i$, ± 1, and they must not satisfy the relation $p^2 + q^2 = 0$.

4. *The General Case.*—The equations (9) are now equivalent to

$$q(p^2 - 1)M = p(p^2 + 1)N, \qquad p(q^2 + 1)M = q(p^2 + 1)N. \tag{13}$$

The determinant of these equations in M and N is $-(p^2 + 1)(p^2 + q^2)$. Since this cannot be zero (as we have already treated these various cases), it follows that both M and N are zero. Hence from equation (8), we find

$$p = \frac{2xz}{x^2 + y^2 - z^2}, \qquad q = \frac{2yz}{x^2 + y^2 - z^2}. \tag{14}$$

The Pfaffian, $dz = pdx + qdy$, may be written in the form

$$2xzdx + 2yzdy - (x^2 + y^2 - z^2)dz = 0. \tag{15}$$

This can be made exact by dividing by $z(x^2 + y^2 + z^2)$. The complete integral is

$$x^2 + y^2 + z^2 = \text{const. } z. \tag{16}$$

These are spheres passing through the origin and with axis as the z-axis.

This completes the proof of our converse of the theorem of Ptolemy on stereographic projection. We have shown that the sphere is the only surface admitting a perspective conformality.

Elsewhere we shall discuss the possibility of two surfaces (neither one plane) admitting conformal perspectivity, a much more difficult problem.

5. *Other Characterizations by Scale Curves.*—The authors have given other characteristic properties of this stereographic map by means of scale curves. These are defined by the $\sigma = dS/ds =$ const. In any conformal

map upon a plane, there are ∞^1 scale curves. It develops that the most famous maps, namely, Mercator projection, stereographic (Ptolemy) and Lambert projections, have certain common characteristic properties.

The only conformal maps of a sphere upon a plane with isothermal scales are the Mercator projections (with straight scales), and the Ptolemy stereographic and Lambert projections (with circular scales), followed by a similitude in the given plane.

If ∞^1 geodesics of the sphere are mapped conformally upon a plane such that they are represented by straight lines, then the maps are stereographic (Ptolemy), or Lambert projections (with circular scales) or Mercator projection (with straight scales). There does not exist a conformal map such that all the ∞^2 great circles are mapped into straight lines of the plane.

The only map with straight scales is the Mercator projection whereas the only maps with circular scales are the stereographic and Lambert maps. Mercator's cartogram straightens out the loxodromes. This is possible not only for the sphere but also for any surface of revolution. We have shown that no further extension is possible.

[1] Kasner and De Cicco, "Ovals Should Be Used to Map Airplane Ranges," *Science News Letter*, 200 (March 25, 1944).

[2] Kasner and De Cicco, "Scale Curves in Conformal Maps," these PROCEEDINGS, 30, 162–164 (1944).

[3] Kasner and De Cicco, "Scale Curves in General Cartography," *Ibid.*, 30, 211–215 (1944).

[4] Kasner and De Cicco, "Geometry of Scale Curves in Conformal Maps," *Am. Jour. Math.*, 67, 157–166 (1945).

[5] De Cicco, "Conformal Maps of Surfaces with Isothermal Systems of Scale Curves," *Ibid.*, 67 (1945).

[6] In our preceding paper, "The Laplace Equation in Space," published in the August PROCEEDINGS, theorems 4 and 5 should be ammended by replacing the Liouville inversion group by the similitude group. Thus while inversion in the plane preserves isothermal families of curves, inversion in space does not preserve isothermal families of surfaces. Conformality is necessary but not sufficient in three (or more) dimensions. This apparently has not been noticed before.

[7] Santola, "Un Theorema sobra Representacion Conforme," *Boletin Matematica Rosario*, 29–40 (1945). Many of the results of this interesting paper were given by Kasner in previous papers on differential elements (1906) and the invarient theory of horn angles (1912–1943), and in the papers by Kasner and De Cicco on cartography.

PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 31 November 15, 1945 Number 11

A NEW EYE COLOR MUTANT IN THE MOUSE WITH ASYMMETRICAL EXPRESSION

BY L. C. DUNN

DEPARTMENT OF ZOÖLOGY, COLUMBIA UNIVERSITY

Communicated October 5, 1945

Most hereditary variations in the paired structures of animals with bilateral symmetry express themselves equally on both sides of the body. Exceptions to this rule are of interest as potential material for revealing new relationships in the mechanism of development. One such exception has been found recently in which an eye color mutation in the mouse frequently expresses itself differently in the right and left eyes.

The observations began while we were studying a case of *heterochromia iridis* which had been noted in a single animal in our stocks. This animal, a brown agouti pied female with a normal brown left eye and a much lighter rose- or ruby-colored right eye proved to be a somatic mosaic with no demonstrable change in the germ cells since we found no eye color variations in 6 F_1 and 31 F_2 offspring from this female. In searching for other possible occurrences of this type of variation we found that Dr. C. H. Danforth of Stanford University was maintaining a stock of silver mice with *heterochromia iridis*, a frequent condition being one eye ruby, the other pink. Dr. Danforth kindly sent us several animals from this stock which we tested by crossing with our heterochromic female. The F_1 consisted of normal black-eyed animals only, indicating that the two occurrences of heterochromia were not due to the same germinal change. In F_2, some animals with ruby eyes like those in the Danforth parent stock appeared, indicating segregation of a recessive ruby mutation.

We then turned our attention to the ruby mutation which had not been described previously. It proved to be a simple recessive to normal, outcrosses to wild type producing only wild type eye color in F_1 and 179 wild type (dark-eyed) and 54 ruby-eyed in F_2. The new eye color was easily scored at birth, the eyes of newborn mutants lacking the iris pigment ring of the wild type and resembling the unpigmented eyes of the mutants pink, pallid and albino.

The F_2 mutants obtained above contained the ruby gene (r) in combination with agouti and with black. Black agouti-ruby superficially resembles brown agouti since the black base of the fur is reduced to dark sepia while the yellow tip is only slightly reduced in intensity. Black-ruby is dull dark sepia or dark slate. Ruby obviously affects black hair pigments in the same direction but to a lesser extent than the mutants pink and pallid, while yellow which is unaffected by the latter mutants is somewhat reduced by ruby. Of the F_2 ruby animals nearly all showed ruby on both sides, but a few showed the heterochromia of the ruby grandparent.

These effects all resemble those of a mutation reported by Sô and Imai.[1] According to Grüneberg[2] (p. 42), who has had the Japanese text of this paper translated, the evidence shows that the ruby mutation of Sô and Imai was an allele of pink, the order of dominance being $P \gg p^r > p$. I have also examined a translation of this paper* and agree with Grüneberg that the evidence for allelism is conclusive. Consequently our ruby was tested by a pink-eyed agouti stock. F_1 was normal dark-eyed; F_2 consisted of 56 normal, 24 ruby-eyed and 28 pink-eyed. This is a sufficiently close approach to 9:3:4 ratio to suggest that our ruby and pink assort independently, the double recessive $pp\ rr$ being pink-eyed. A number of F_2 pink-eyed animals were tested. While most of them proved to be $pp\ Rr$, a few were $pp\ rr$. A double recessive stock of pink-eyed animals $pp\ rr$ was prepared of which the members were indistinguishable in eye color or coat color from the normal pink-eyed stock. Linkage relations of the new ruby with pink were tested by matings $\frac{PR}{pr} \times pp\ rr$ and $\frac{Pr}{pR} \times pp\ rr$. The former produced 36 dark-eyed (PR) and 87 light-eyed (pR, Pr and pr) young; the latter, 35 dark-eyed and 91 light-eyed. Combining these, the recombinations, 128, constituted almost exactly half of the 249 gametes tested, indicating free assortment of p and r.

Since two phenotypically indistinguishable but independent pink-eye mutations have been recorded in the mouse, it occurred to us that the Japanese investigators may have employed the p^2 (pallid) allele in their tests. We therefore tested ruby by p^2.† This produced only wild type offspring (38 observed) indicating that r is not a mutation at the p^2 locus. Similar tests by albino likewise gave no evidence of allelism of r and c^a. We may therefore assume that r marks a new locus in the mouse.

For reasons of interest in this laboratory we tested r for location in the T (Brachyury) chromosome. The tests produced the following results: $\frac{+T}{r+} \times ++$ gave 100 $++$, 114 $+\ T$, 124 $r+$, 100 $r\ T$. This fits an assumption of independent assortment ($p = 0.3$) and indicates that the r locus is probably not in the T chromosome.

Returning now to the asymmetrical expression of the new mutant we have noted that in the original stock as received from Professor Danforth, 53 animals carefully examined showed the following eye color distribution, as compared with rubies extracted from outcrosses to non-ruby stocks:

	BOTH EYES RUBY	LEFT EYE RUBY RIGHT EYE PINK	LEFT EYE PINK RIGHT EYE RUBY	BOTH EYES PINK
Original stock	22	11 + 5[a]	5 + 5[b]	5
Extracted ruby	50	4	3	1

[a] These were recorded as left eye dark ruby, right eye light ruby.
[b] These were recorded as right eye dark ruby, left eye light ruby.

While it is clear from these observations that asymmetry in eye color is frequent in the original stock (26 asymmetrical out of 53 carefully observed for this character), it is not evident that the asymmetry is due to the ruby gene, since after extraction from outcrosses the same ruby gene showed only about 12 per cent (7/57) asymmetry. The original stock was silver black pied (*aa BB sisi ss*). After outcrossing this to black agouti (*AA BB SiSi SS*), black and tan (*a^t a^t BB SiSi SS*) and pink-eyed black stocks (*aa BB pp SiSi SS*) the asymmetry of extracted ruby-eyed animals showed no relation to *A*, a^t or *a*, nor to the *Si-si* alleles. However out of 7 asymmetrical rubies, 5 were pied. Since the total number of pied animals observed in segregating generations was 9, we have proportions of asymmetrical ruby-eyed as 5/9 among pied and 2/48 among non-pied. The difference between these proportions is highly significant ($p = < 0.001$). It is probable therefore that the asymmetrical expression of ruby is greatly enhanced by pied spotting. This is to be taken as a preliminary indication only, since extended and systematic studies of other factors influencing asymmetrical expression have not been made.

In the present data there is some tendency for the left eye to be darker than the right, but the observations are too few to provide a test of this point. It is apparent that the expression of the ruby mutation depends upon an equilibrium which is influenced by both genetic and developmental factors which deserve further study.

The ruby mutation of Sô and Imai was similarly quite variable in density of eye color and this appeared to depend on other (unanalyzed) genetic factors. The Japanese authors also noted *heterochromia iridis* frequently in the ruby-eyed stock. They believed it to be genetically influenced but because of the weakness of such animals were unable to analyze it.

The mutation of Sô and Imai and that described here are the only two in mice in which heterochromia has been observed. This fact and their other similarities render them phenotypically almost identical yet they appear to be mutations at different loci. The only alternative to this is to assume

that the pink mutation with which Sô and Imai's ruby proved to be allelic was a third mutation, different from the p and p^2 of American and European stocks. This is unlikely since the original pink mutation was introduced into Europe from Japan.

* Made by Toshi Miyazaki of Columbia University.

† Kindly supplied by Jackson Memorial Laboratory.

[1] Sô, M., and Imai, Y., "On the Inheritance of Ruby Eye in Mice," *Japanese Jour. Genet.*, 4, 8–9 (1926).

[2] Grüneberg, H., *The Genetics of the Mouse*, University Press, Cambridge, 1943, 412 pages.

ON NON-LINEAR PHENOMENON OF SELF-ROLLING

BY N. MINORSKY

DAVID TAYLOR MODEL BASIN, WASHINGTON, D. C.

Communicated October 4, 1945

Recently developed methods of non-linear mechanics offer interesting applications to a number of problems which have been known for a long time but which have remained unexplored so far. A typical problem of this kind is the so-called *self-rolling* of a ship provided with an antirolling stabilizing system. It is well known that, if the phase of the control action is reversed, the ship, originally at rest in still water, begins to roll with gradually increasing roll angles until the amplitude reaches a stationary value after which the self-rolling continues indefinitely. Although the physical nature of this phenomenon is sufficiently clear, it is impossible to determine the stationary amplitude from the linear theory for what eventually limits the amplitude is precisely the action of the various non-linearities which are generally left out in a simplified linear theory.

The differential equation of rolling in still water is

$$I\ddot{\theta} + a_1'\dot{\theta} + a_2'\dot{\theta}^2 + Dh(\theta) = 0 \tag{1}$$

where θ is the angle of roll, D is the displacement of the ship, I is the virtual moment of inertia about the longitudinal axis through the center of gravity, a_1' and a_2' are the so-called Froude's coefficients of resistance to rolling (bilge keels, skin friction), and $h(\theta)$ is the lever arm of the restoring couple. For small angles of roll $h(\theta) \cong h_1\,\theta$ where h_1 is the initial metacentric height assumed to be constant. For larger angles one can approximate the quantity $Dh(\theta)$ by an expression of the form

$$Dh(\theta) = c_1'\theta - c_3'\theta^3 \tag{2}$$

which means that the restoring couple decreases for large θ. The coefficients c_1' and c_3' depend on the form of the hull and can be computed for a given ship.

The stabilizing moment S in general is a certain function of $\dot{\theta}$. In practice one endeavors to obtain a linear law $S = b\dot{\theta}$, where b is a constant. In reality owing to the usual form of the characteristic of the stabilizing power plant it is found[1] that the stabilizing moment for larger $\dot{\theta}$ increases less rapidly than according to the above simple relation and can, therefore, be approximated also by an expression of the form

$$S = b_1'\dot{\theta} - b_3'\dot{\theta}^3. \qquad (3)$$

The coefficients b_1' and b_3' can be determined from the characteristic of the equipment. Since we are interested here in self-rolling, we have to add $-S$ on the left side of equation (1). The differential equation of self-rolling will then be

$$I\ddot{\theta} + a_1'\dot{\theta} + a_2'\dot{\theta}^2 - b_1'\dot{\theta} + b_3'\dot{\theta}^3 + c_1'\theta - c_3'\theta^3 = 0. \qquad (4)$$

Dividing by I and setting $c_1'/I = \omega_1^2$; $a_1'/I = a_1$; $a_2'/I = a_2$, etc., this equation becomes

$$\ddot{\theta} + \omega_1^2\theta = (b_1 - a_1)\dot{\theta} - a_2\dot{\theta}^2 - b_3\dot{\theta}^3 + c_3\theta^3 = f(\theta, \dot{\theta}). \qquad (5)$$

This is the non-linear differential equation of our problem. We will assume that the upper bound of the non-linear function $f(\theta, \dot{\theta})$ is small so that the non-linear motion specified by equation (5) remains in the neighborhood of the linear motion corresponding to $f(\theta, \dot{\theta}) = 0$. This assumption is justified in practice since it is observed that the self-excited stationary oscillation is practically sinusoidal. Owing to this assumption it is possible to apply the so-called method of the first approximation[2] and solve the problem.

We shall leave out the case when $b_1 - a_1 = \alpha < 0$ as no self-rolling is possible in this case and will consider only the case when $\alpha > 0$, that is, when the linear term $b_1\dot{\theta}$ due to the stabilizer outweighs the linear term $a_1\dot{\theta}$ of the Froude resistance to rolling. One ascertains this from the theory of Poincaré,[3] viz.: $\alpha = 0$ is a bifurcation point; for $\alpha < 0$ the singular point of equation (5) is a stable focal point; for $\alpha > 0$ it is an unstable focal point which means that the amplitudes begin to increase from rest.

In order to obtain the stationary amplitude $\bar{\theta}_0$ we look for a solution of the form $\theta = \theta_0 \sin\psi$, where θ_0 is an amplitude (not necessarily stationary) and $\psi = \omega_0 t + \varphi$ is the so-called *total phase*. The quantities θ_0 and ψ are determined by the differential equations

$$\frac{d\theta_0}{dt} = \frac{1}{2\pi\omega_0}\int_0^{2\pi} f(\theta_0 \sin\lambda, \theta_0\omega_0 \cos\lambda) \cos\lambda d\lambda = \frac{1}{2\pi\omega_0}\phi(\theta_0) \qquad (6)$$

$$\frac{d\psi}{dt} = \Omega(\theta_0) = \omega_0 - \frac{1}{2\pi\theta_0\omega_0}\int_0^{2\pi} f(\theta_0 \sin\lambda, \theta_0\omega_0\cos\lambda)\sin\lambda d\lambda$$
$$= \omega_0 - \frac{1}{2\pi\theta_0\omega_0}\Psi(\theta_0). \quad (7)$$

The condition for a stationary amplitude is

$$\phi(\bar{\theta}_0) = 0. \quad (8)$$

In carrying out these calculations attention must be paid to the term $a_2\dot{\theta}^2$ in equation (5). In the integration between 0 and 2π, this term, for obvious physical reasons must be written as $a_2|\dot{\theta}|\dot{\theta}$. When all integrations are made, condition (8), upon the cancellation of the unstable solution $\bar{\theta}_0 = 0$ reduces to the following

$${}^3/_4 b_3\pi\omega_0^3\bar{\theta}_0^2 + {}^8/_3 a_2\omega_0^2\bar{\theta}_0 - \alpha\omega_0\pi = 0. \quad (9)$$

The stationary amplitude $\bar{\theta}_0$ is given by the positive root of this equation, viz.:

$$\bar{\theta}_0 = \frac{16}{9\pi\omega_0}\frac{a_2}{b_3}\left[\sqrt{1 + \frac{27}{64}\pi^2\frac{\alpha b_3}{a^2_2}} - 1\right]. \quad (10)$$

From this expression it follows that for $\alpha < 0$ there exists no positive value for $\bar{\theta}_0$ which means that the self excitation is impossible. This follows also from the theory of Poincaré as was just mentioned. For very small positive value of α we have

$$\bar{\theta}_0 \cong \frac{3}{8}\frac{\pi\alpha}{\omega_0 a_2}. \quad (11)$$

Hence, for a relatively small intensity of the self-rolling control, the stationary amplitude increases in proportion to α. If α remains fixed in that range, the stationary amplitude is in an inverse proportion to the coefficient a_2. The frequency correction ω_1 is obtained from equation (7). Carrying out the indicated integrations one obtains

$$\Omega(\bar{\theta}_0) = \omega_0 - \omega_1 = \omega_0 - \frac{3c_3}{8\omega_0}\bar{\theta}_0^2 \quad (12)$$

where $\bar{\theta}_0$ is given by equations (10) or (11). It is seen thus that the non-linear stationary frequency is somewhat lower than the corresponding linear undamped frequency ω_0. The frequency correction ω_1 depends on $\bar{\theta}_0$.

If the non-linearities (a_2 for the ship and b_3 for the stabilizer) are known, equation (10) permits determining $\bar{\theta}_0$. If, however, a_2 is unknown and the stabilizing equipment is available, by observing $\bar{\theta}_0$ and computing b_3 it is possible to determine the Froude constant a_2.

[1] Minorsky, N., "On Mechanical Self-Excited Oscillations," *Proc. Nat. Acad. Sci.*, **30**, 308–314 (1944).

[2] Kryloff, N., and Bogoliuboff, N., "Introduction à la Mécanique non linéaire," Kieff (1937); also free English translation of S. Lefschetz, Princeton University Press (1943).

[3] Poincaré, H., "*Oeuvres*," Paris, Volume 1; also D. Taylor Model Basin report No. 534 (1944).

---

*EFFECTS OF EXPOSURE TO ULTRA-VIOLET LIGHT ON SUBSEQUENT DARK ADAPTATION**

BY ERNST WOLF

BIOLOGICAL LABORATORIES, HARVARD UNIVERSITY

Communicated October 15, 1945

It has been shown that ultra-violet light below 365 mμ affects the visual mechanism in baby chicks by raising the threshold for recognition of flicker. For the return of the threshold to its normal level many hours are required.[1]

For the human eye it is known that the course of dark adaptation can be modified by previous exposure to different levels of intensity[2] and to selected portions of the visible spectrum of the same photic intensity.[3] The influence of the ultra-violet component of light has thus far not been taken into consideration. With data on effects of ultra-violet on the chick's eye available, its dark adaptation was studied after previous exposure to lights of varying ultra-violet content.

Preceding dark adaptation measurements, chicks 4 to 12 days old are exposed singly to three 250 watt mercury vapor lamps (GE type H-5), while confined to a cylindrical wire mesh cage, but free to move with access to food and water. The lamps are mounted at eye level, equally spaced outside the cage and surrounded by a reflecting surface. The illumination at the center of the cage is 880 millilamberts as measured with a Weston photoelectric photometer. Lamp housings permit the insertion of filters for selection of specific spectral regions. For exposure to the light a standard period of 10 minutes is chosen.

With the completion of light exposure the chick is transferred rapidly into a cylindrical glass cage and brought into the apparatus for testing responses to flicker.[4] Flicker is produced by a cylindrical system of equally spaced dark vertical bars, rotating around the animal. At threshold level of light intensity the chick responds with a movement of its head, following the direction of the stripe motion through an angle of 45 to 90° and returning abruptly to its initial position (head nystagmus). In determining the level of dark adaptation at any given moment it is important that the width of the bars be big enough to be seen at minimum intensities.

In our case the visual angle of each bar measures 36° at the center of the glass cage. Flicker recognition also depends upon the velocity of passage of the bars in front of the eye and must be slow at threshold intensities. The speed chosen is therefore the slowest possible at which precise responses occur, and this is about 1 to 1.5 flicker per second. The first test is made one minute after cessation of illumination and the threshold intensity for response recorded. Then the chick remains in complete darkness and the tests are repeated every minute up to 10, then every two

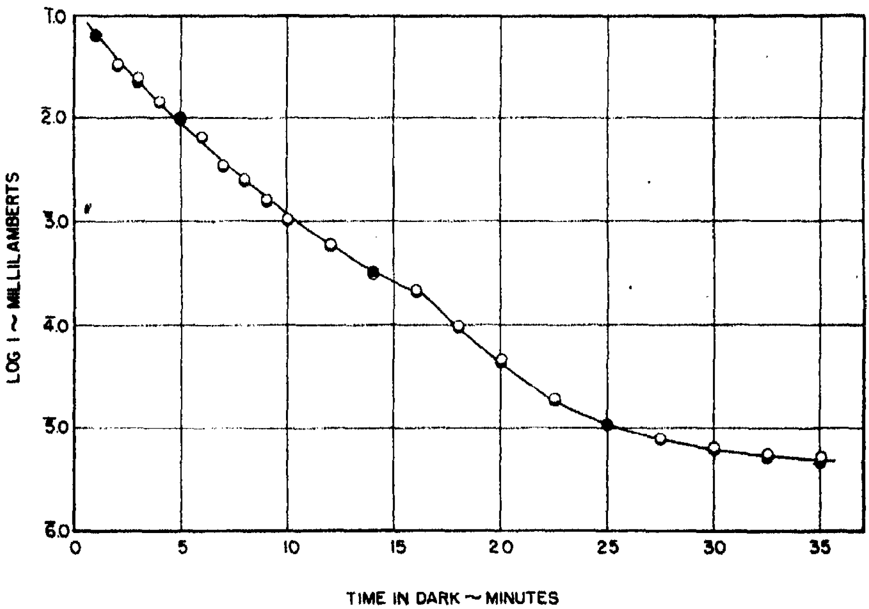


FIGURE 1

Relations of threshold intensities of illumination and time of dark adaptation for the eye of the baby chick, after exposure to white light (880 millilamberts) for 10 minutes. Two series of measurements, each containing six individuals, are represented by different symbols.

minutes up to 20, and every 2.5 minutes up to 35 minutes of dark adaptation. For each dark adaptation curve plotted the averaged readings of six individuals are taken to smooth out individual deviations.

Detrimental effects of ultra-violet light on visual threshold, as described previously,[1] showed that wave-lengths below 365 mμ raise the threshold, whereas longer waves have no effect. For a normal dark adaptation curve it is therefore necessary to filter out from the exposure light any portion of the ultra-violet that might have an influence upon its course. A filter with the designation XY 91 VO of the Polaroid Corporation which trans-

mits only wave-lengths above 400 mμ and does not produce any noticeable changes of hue in the visible was selected to provide the exposure light for our standard curve.

In figure 1 the mean values of threshold intensities for response are represented. The curve has the general form of a dark adaptation curve typical for a visually duplex vertebrate eye (cf. human eye). In darkness the sensitivity increases rapidly in the beginning, then more slowly, until it reaches a relatively steady level after about 35 minutes, when the experiment is broken off. Due to the presence of rods and cones in the retina of the chick's eye, the curve shows a break dividing the process of adaptation into a cone and a rod component. The break, sharply indicated in the graph, because of the fact that the point at 16 minutes may equally well be fitted to either segment, would not be so sharp, if all individual readings were plotted; the transition would rather be rounded off. The fact that the break occurs after about 16 minutes of dark adaptation is of interest in so far as for the human eye (large visual field and white exposure light) the transition occurs much earlier.[5] There is also a considerable difference in relative extent of the cone and rod range. For the human eye, after light adaptation to intensities of the same order of magnitude, the cone range covers 1.5–2 log. units, the rod range 3 log. units. In the chick the cone range is 3 log. units, whereas the rod range is only 1.5. The relative participation of the two groups of receptors is roughly the reverse in the chick from that in man. These facts are in accordance with morphological differences between the human and the chick's eye. The retina of the chick, as in most diurnal birds, is predominantly a cone retina with the interspersion of relatively few rods, whereas the human retina has, besides the high concentration of cones in the fovea, relatively few cones distributed among the peripheral rods.[6]

Knowing the course and extent of dark adaptation of the chick's eye after exposure to ordinary light free of ultra-violet, we are in a position to add to the exposure light certain ranges of the ultra-violet spectrum by the use of different absorption filters. The range between 300 and 400 mμ is of particular interest. A series of filters with cut-offs at different wave-lengths in this region provided previously a differential influence[1] and is used here for a study of similar effects on the course of dark adaptation. The filters are in their order of transmission: Corning 774 (heat-resisting clear chemical glass), AO crown glass 1045, ordinary plate glass, AO Cruxite 1794, Corning 3850 (greenish Nultra), AO Calobar C-1827, AO 2614, Corning 3389 (Noviol shade A), XY 91 VO of the Polaroid Corporation and Corning 9863 (red-purple Corex A). The transmissions of these filters, determined with a Beckman spectrophotometer, are given in figure 2.

With Corning 774 (jacket of the GE H-5 lamp as furnished by the manufacturer) in front of the mercury vapor arc during light exposure, it is

found that in subsequent dark adaptation tests for the first 16 minutes the course is identical with the standard curve given in figure 1. From the point of cone-rod transition on, where we would normally expect a considerable drop in threshold intensities, only a slow decline is apparent and we arrive very soon at a steady level so that after 35 minutes in the dark it is about 1.3 log. units higher than the normal. Due to the presence of the ultra-violet in the exposure light, the state of dark adaptation is therefore only one-twelfth of the normal at the termination of the tests.

A gradual reduction of the extent of the ultra-violet band from 290 to 360 mμ by crown, plate glass, Cruxite and Corning 3850 affects the course and final threshold of dark adaptation in correspondence to the elimination of the shorter wave-lengths. At 35 minutes the difference in threshold intensity from the normal is for crown about 1.1, for plate glass 0.6, for

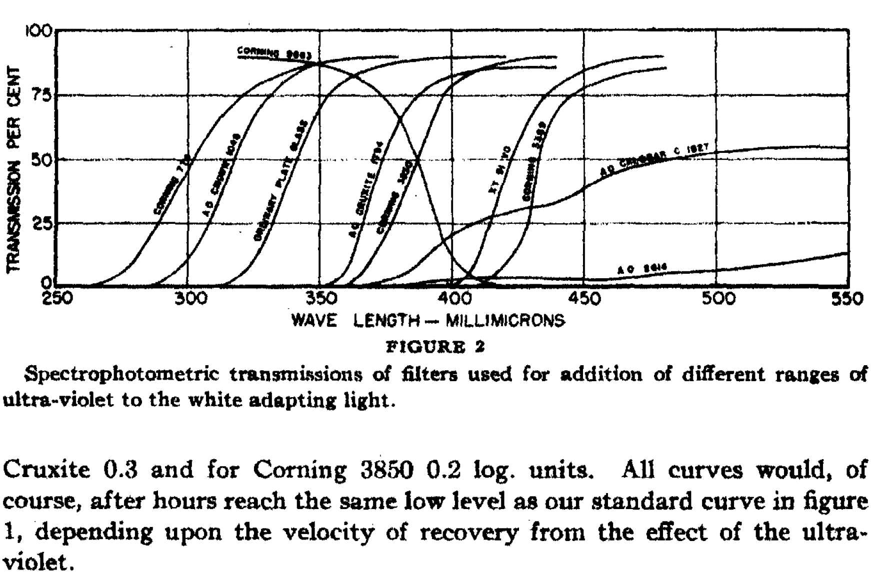


FIGURE 2

Spectrophotometric transmissions of filters used for addition of different ranges of ultra-violet to the white adapting light.

Cruxite 0.3 and for Corning 3850 0.2 log. units. All curves would, of course, after hours reach the same low level as our standard curve in figure 1, depending upon the velocity of recovery from the effect of the ultra-violet.

Exposure light which does not contain any wave-lengths shorter than 365 mμ results in dark adaptation curves which are identical with our standard curve. Corning 3389 has a cut-off at higher wave-lengths than XY 91 VO, and the experimental readings are found practically identical for both. AO Calobar which transmits about 1.5 per cent at 365 mμ and AO 2614 which begins to transmit at 380 mμ are both tinted glasses and have therefore a lower transmission in the visible compared with the colorless filters thus far considered. For Calobar the maximum transmission at 535 mμ is 55 per cent, for AO 2614 at 625 mμ about 20 per cent. Both reduce therefore the brightness of the exposure light. The effect of this decrease is noticeable when following the individual points along the curve;

they are consistently on the low side. The final thresholds for both filters fall closely within the scatter range of the normal curve. It may therefore

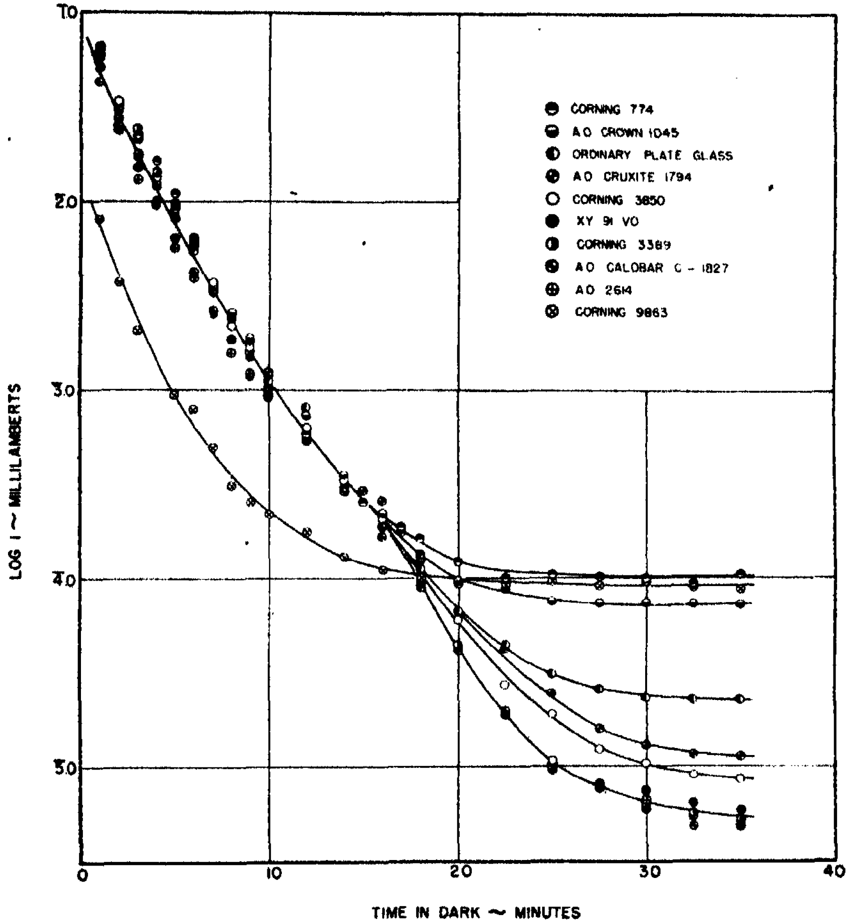


FIGURE 3

The course of dark adaptation of the eye of the baby chick, measured after preceding exposure to white light containing different ranges of the ultra-violet spectrum. In proportion to the increase of the ultra-violet component below 365 mμ the final dark adaptation thresholds are raised to higher intensities, indicating a reduction in potential dark adaptation by the ultra-violet. Light containing wave-lengths above 365 mμ does not produce such effects.

be assumed that any ultra-violet above 365 mμ has no effect upon the course of dark adaptation.

Of particular interest are the results with Corning filter 9863 in front of the exposure light. This filter transmits only a small portion of the far red,

a small amount of violet, but almost as much of the effective ultra-violet as Corning 774. During exposure to light the chick receives a dim purple light which causes adaptation to a lower brightness. In subsequent tests this becomes apparent in the fact that the first threshold readings fall about 0.9 log. units below those with white light. And since the ultra-violet component transmitted by this filter is almost equal to that of Corning 774, the ensuing curve has a shape quite different from all others, so that its final threshold values are nearly as high above the normal as for Corning 774. An equivalent ultra-violet effect is therefore obtained under exclusion of the visible.

The data are presented in figure 3. Each symbol represents the mean of readings on six individuals in each series. It should be noted that for better separation of the curves in this graph the ordinates are twice as high as in figure 1.

Considering the change in the course of dark adaptation due to the presence of different amounts of ultra-violet in the exposure light, it is noteworthy that the cone segment of the curve is unaffected, while beyond the transition to the rod segment the ultra-violet effect becomes apparent in varying degree. An analysis of the retinal pigments of the chick shows that the cones possess highly absorbing pigments, whereas the rods are pigment-free.[7] On this basis it may be assumed that the ultra-violet acts primarily on the rods, whereas the cones are protected by their pigments, and hence the pronounced effect on the rod segment of the dark adaptation curve.

Since no data on the influence of ultra-violet light upon the dark adaptation of the human eye are available at present, we might on the basis of physiological similarities of vertebrate eyes assume that similar differences in final dark adaptation thresholds should be expected in human observers, depending upon their exposure to ultra-violet previous to test. It should not be surprising to find differences among subjects of different occupations, indoors or outdoors, or at various latitudes and elevations, particularly if we assume, as we probably may, that the action of ultra-violet is cummulative.

Summary.—The course of dark adaptation in baby chicks is studied after exposure to white adapting light containing different amounts of ultra-violet radiation between 280 and 400 mμ. Exposure light of wavelengths longer than 365 mμ results in uniform dark adaptation curves. Any addition of ultra-violet below 365 mμ prevents complete adaptation, raising the final threshold considerably above the normal level. The difference between complete and partial dark adaptation due to exposure to the short waves is proportional to the extent of the ultra-violet spectrum.

* This research was supported by a grant of the American Optical Company.

[1] Wolf, E., these PROCEEDINGS, 31, 236 (1945).

[2] Winsor, C. P., and Clark, A. B., *Ibid.*, 22, 400 (1936); Hecht, S., Haig, Ch., and Chase, A. M., *Jour. Gen. Physiol.*, 20, 831 (1937); Wald, G., and Clark, A. B., *Ibid.*, 21, 93 (1937); Hecht, S., these PROCEEDINGS, 23, 227 (1937).

[3] Rowland, W. M., and Sloan, L. L., *Jour. Opt. Soc. Am.*, 34, 601 (1944); Hecht, S., and Yun Hsia, *Ibid.*, 35, 261 (1945).

[4] Wolf, E., and Zerrahn-Wolf, G., *Jour. Gen. Physiol.*, 19, 495 (1936); 19, 229 (1935); Crozier, W. J., and Wolf, E., *Ibid.*, 23, 531 (1940).

[5] Hecht, S., *Ibid.*, 4, 113 (1921); Hecht, S., "The Nature of the Photoreceptor Process," Handbook of Gen. Exp. Psychology, Clark University Press, Worcester, 1934. Hecht, S., Haig, Ch., and Wald, G., *Jour. Gen. Physiol.*, 19, 321 (1935).

[6] Schultze, M., *Arch. f. mikr. Anat.*, 2, 175 (1866); Walls, G., *The Vertebrate Eye*, Cranbrook Inst., Bloomfield Hills, Michigan, 1942.

[7] Wald, G., and Zussman, H., *Jour. Biol. Chem.*, 122, 449 (1938).

CARBON DIOXIDE UTILIZATION IN THE SYNTHESIS OF ACETIC AND BUTYRIC ACIDS BY BUTYRIBACTERIUM RETTGERI

BY H. A. BARKER, M. D. KAMEN* AND VICTORIA HAAS

DIVISION OF PLANT NUTRITION, UNIVERSITY OF CALIFORNIA, BERKELEY

Communicated October 9, 1945

Butyribacterium rettgeri causes a modified butyric acid type fermentation of lactate, the main products being carbon dioxide and acetic and butyric acids.[1] This fermentation is remarkable because of the low yield of carbon dioxide and the high yield of fatty acids. In the usual type of butyric acid fermentation, such as is carried out by *Clostridium saccharobutyricum*,[2] for example, not less than one mole of carbon dioxide and not more than one mole of C_2 derivatives (acetic acid, butyric acid $\times$ 2, etc.) are formed per mole of triose. *B. rettgeri*, however, produces only 0.4 mole of carbon dioxide and up to 1.2 moles of C_2 derivatives per mole of lactate. These unusual yields suggest that this organism utilizes part of the carbon dioxide produced from lactate for the synthesis of fatty acids. This possibility has been investigated using the long-lived radioactive carbon isotope, C^{14}, as a tracer to follow the transformations of carbon dioxide, with the results reported below.

Experimental.—To investigate the possible conversion of carbon dioxide into fatty acids, *B. rettgeri* was allowed to grow in a medium containing lactate and C^*O_2. (C^* indicates carbon labeled with C^{14}.) The substrate used and the products formed were estimated quantitatively and the C^{14} content of each compound was determined.

A basal medium of the following composition in g. per 100 ml. was used: sodium lactate 0.5–2.0, Difco yeast extract 0.3, yeast autolysate 0.3,

$(NH_4)_2SO_4$ 0.05, K_2HPO_4 0.1, KH_2PO_4 0.4, $MgSO_4.7H_2O$ 0.02, cysteine hydrochloride 0.05, $FeSO_4.7H_2O$ 0.0005, pH 6.2. A sterile solution of labeled sodium carbonate was added to the basal medium after autoclaving to give a final pH of 7.0–7.6. After inoculating with *B. rettgeri*, strain 32, samples were removed to determine the initial lactate and carbon dioxide. An Oxsorbent seal was then applied to remove oxygen. Cultures were incubated at 30–35°C. until growth ceased and the products of the fermentation were estimated quantitatively, separated and their C^{14} contents determined. The separation of the fatty acids was accomplished by the distillation procedure of Schicktanz, *et al.*[3] The method of determining C^{14} activity has already been described.[4]

TABLE 1

THE FERMENTATION OF LACTATE IN THE PRESENCE OF C^*O_2

(Experiment 1)

COMPOUND	MM/10 ML.	CTS./MIN./MM.	TOTAL CTS./MIN.
Lactic acid, decomposed	0.463	...	...
Carbon dioxide, initial	0.115	33,800	3,890
Carbon dioxide, final	0.298	4,800	1,430
Acetic acid	0.252	2,470	620
Butyric acid	0.177	6,350	1,120

The results of a typical experiment are given in table 1. It can be seen that approximately 63% of the C^{14} was lost from the carbon dioxide during the fermentation of 0.46 mM of lactate per 10 ml. Of the total C^{14} lost, about 25% and 46% was recovered as acetic and butyric acids, respectively. In another experiment the corresponding values were 26% and 53%, while, in addition, 1% of the C^{14} was recovered as caproic acid, 7% as trichloracetic acid insoluble cell material and 3% as unidentified nonvolatile compounds. The total recovery of transformed C^{14} varied from 80 to 90%.

The distribution of C^{14} in the acetic acid was determined by the decarboxylation of the barium salt as previously described[4] after the purity of the product was established by a Duclaux distillation. The data given in table 2 show that both methyl and carboxyl groups contain the isotope,

TABLE 2

DISTRIBUTION OF C^{14} IN ACETIC ACID

(The figures give the percentage of the total C^{14} in each atom)

CARBON ATOM	EXPT. 2	EXPT. 3
Carboxyl	56.5 ± 2	57 ± 1
Methyl	43.5 ± 2	43 ± 1

though the amount is not the same in the two positions. The carboxyl group contains significantly more C^{14} than the methyl group.

The approximate distribution of C^{14} in butyric acid was determined by a combination of two methods: (1) the oxidation of the ammonium salt with hydrogen peroxide as described by Wood, *et al.*,[5] and (2) the decarboxylation of barium butyrate.

Wood, *et al.*, have shown that the acetone formed in the H_2O_2 oxidation of butyric acid is derived from the alpha, beta and gamma carbon atoms. By further oxidizing the acetone with alkaline iodine solution, iodoform and acetic acid are formed. The iodoform originates from the alpha and gamma positions in butyric acid and may be used to determine the average activity of these two carbon atoms. The acetic acid originates from the alpha or gamma and beta positions. When the molar activities of the iodoform and acetic acid are known, the activity of the carboxyl carbon in acetic acid, which is equivalent to the beta position in butyric acid, can be calculated by difference.

Decarboxylation of barium butyrate was used to determine the activity of the carboxyl carbon. The reaction was carried out at about 400°C. Control experiments with carboxyl-labeled butyric acid, prepared by Grignard synthesis,[6] showed that approximately half (actually 52%) of the carboxyl activity appeared in the barium carbonate formed by decarboxylation. The reaction is therefore reliable for quantitatively determining the amount of C^{14} in the carboxyl group of the fermentation butyric acid.

TABLE 3

DISTRIBUTION OF C^{14} IN BUTYRIC ACID

(Experiment 3)

CARBON ATOM	PERCENTAGE OF TOTAL C^{14}
Carboxyl	39.3
Alpha and gamma (average)	30.8
Beta	29.9

The results presented in table 3 demonstrate that all four positions in butyric acid contain carbon derived from carbon dioxide. As in acetic acid, the distribution of C^{14} in butyric acid is not entirely uniform. The carboxyl carbon atom contains significantly more C^{14} than the atoms in other positions. The appearance of more C^{14} in carboxyl carbon relative to alpha and gamma carbon was to be expected, since butyric acid is probably formed (see below) by a condensation of two molecules of acetic acid or derivatives thereof,[5, 7] and the methyl group of the acetic acid from which the alpha and gamma carbon atoms would be derived, contains less C^{14} than the carboxyl group. On the same basis, the beta carbon atom of butyric acid should have the same C^{14} content as the carboxyl group, which is not in accordance with the observed result. A possible explanation for this apparent discrepancy is that the reported C^{14} content of the beta carbon atom may be in error. Due to the fact that this value is calcu-

lated by difference from the data for acetic acid and iodoform derived from the oxidation of acetone, it is probably less reliable than the values for the other carbon atoms.

Having demonstrated the conversion of carbon dioxide to fatty acids and cell material, it seemed desirable to determine the total quantity of carbon dioxide formed per mole of lactate decomposed. It is obvious that the total carbon dioxide production ($CO_{2\ total}$) must exceed the observed or net carbon dioxide production ($CO_{2\ obs.}$) by the amount used for synthetic reactions ($CO_{2\ used}$), i.e.,

$$CO_{2\ total} = CO_{2\ obs.} + CO_{2\ used.} \qquad (1)$$

The carbon dioxide used and therefore the total carbon dioxide produced can be calculated from data, such as is presented in table 1, showing the decrease in C^{14} content of the carbon dioxide during the fermentation of a known amount of lactate, if it is assumed that the formation of carbon dioxide from lactate is the only process causing a dilution of the C^*O_2. More specifically, it must be assumed that no exchange of C^{14} occurs between carbon dioxide and any of its reduction products such as acetic and butyric acids.

The absence of such exchange was verified by an experiment in which lactate was fermented in the presence of acetic acid labeled in both the methyl and carboxyl groups; the carbon dioxide was unlabeled. After the fermentation, no C^{14} (2 ± 3 cts./min.) could be detected in carbon dioxide, while about 57% of the C^{14} was present in butyric acid (445 ± 10 cts./min.). Incidentally, this result also proves that butyric acid is formed from acetic acid.

Another assumption involved in the calculation is that the carbon dioxide inside and outside the cells is in isotopic equilibrium at all times. No direct test of the validity of this assumption has been attempted. If isotopic equilibrium does not exist, the C^{14} content of carbon dioxide inside the actively metabolizing cells will be lower than that outside. Such a condition would result in a lowering of the calculated total carbon dioxide production since the utilization of C^*O_2 and hence its dilution would be less than anticipated.

The calculation of $CO_{2\ used}$ and $CO_{2\ total}$ follows the method developed for the acetic acid fermentation caused by *Clostridium thermoaceticum*[4] with the exception that the quantity V representing the carbon dioxide concentration throughout the fermentation is variable instead of constant. At any instant, more carbon dioxide is being produced from the decomposition of lactate than is being used in synthesis; consequently, a net increase in carbon dioxide concentration is observed. Let x be the quantity of C^*O_2 per unit volume at any time during the fermentation, x_i and x_f denoting the initial and final C^*O_2 concentrations. V will now represent the

total concentration of carbon dioxide ($C^*O_2 + CO_2$) present at any time. For every mole of carbon dioxide observed to be produced ($CO_{2\ obs.}$) as a result of the breakdown of lactate, A moles of $CO_2 + C^*O_2$ are converted to acetic acid. The decrease in C^*O_2 concentration ($-\Delta x$) is given by the expression

$$-\Delta x = (A\,\Delta V/V + \Delta V)\cdot x. \qquad (2)$$

The differential form of this relation

$$-dx/x = A\ dV/V \qquad (2a)$$

is integrated between the limits x_i and x_f for x, and V_i and V_f for V, where V_i and V_f denote the initial and final observed concentrations of carbon dioxide.

Thus

$$-\int_{x_i}^{x_f} dx/x = \ln x_i/x_f = A \int_{V_i}^{V_f} dV/V = A \ln V_f/V_i \qquad (3)$$

or

$$A = \log x_i/x_f/\log V_f/V_i. \qquad (3a)$$

Using the data of table 1 where V_i = 0.115 mM, V_f = 0.298 mM, x_i = 3,890 cts./min., x_f = 1,430 cts./min., and the lactate decomposed is 0.463 mM,

$$A = \log 2.72/\log 2.59 = 0.435/0.413 = 1.05$$

and $CO_{2\ total}$ = 0.395 mM + 0.395·1.05 mM = 0.81 mM/mM lactate. In three other experiments the values of A ranged from 0.91 to 1.02 and the values of $CO_{2\ total}$ from 0.62 to 0.81. The lower values were obtained from experiments in which higher concentrations of lactate were fermented (1.0–1.3 mM/10 ml.). However, the data are too few to permit generalization concerning the relation between the quantity of lactate fermented and the functions in question.

Although the above results demonstrate that much more carbon dioxide is produced than accumulates, nevertheless the calculated total carbon dioxide production appears to be significantly lower than one mole per mole of lactate that would be expected on theoretical grounds. This result may well be due to the fact that the carbon dioxide inside the fermenting cells is not in isotopic equilibrium with the carbon dioxide in the outside medium. It has already been noted that lack of isotopic equilibrium would lower the calculated carbon dioxide utilization and total production below the true values.

From the data obtained in these experiments, the fermentation of lactate by *B. rettgeri* can be fitted into the following simplified reaction scheme if we disregard the minor quantitative discrepancies noted above.

$$CH_3CHOHCOOH + H_2O = CH_3COOH + CO_2 + 4H \quad (4)$$

$$2CO_2 + 8H = CH_3COOH + 2H_2O \quad (5)$$

$$2CH_3COOH + 4H = CH_3CH_2CH_2COOH + 2H_2O \quad (6)$$

Reaction (4) represents the oxidation of lactic acid to approximately one mole each of carbon dioxide and acetic acid; the primary product may well be acetyl phosphate or some other C_2 compound convertible into acetic acid rather than the acid itself. Reaction (5) represents the condensation and reduction of two moles of carbon dioxide to acetic acid; this reaction occurs in such a way that both the methyl and carboxyl groups are derived from carbon dioxide. Reaction (6) represents the condensation and reduction of two moles of acetic acid or some related compound to butyric acid. Neither reaction (5) nor (6) goes to completion since they compete for the available hydrogen from reaction (4). The caproic acid which is produced in small yield may be formed by a further condensation of the same type as reaction (6), possibly involving butyric acid and a C_2 compound.[7]

In conclusion it may be pointed out that *B. rettgeri* is the first non-sporulating bacterium and the fourth anaerobe[4, 8, 9] that has been shown to cause a total synthesis of acetic acid from carbon dioxide.

The authors are indebted to Professor E. O. Lawrence and members of the staff of the Radiation Laboratory, and to Dr. T. H. Norris of the Chemistry Department for facilities used in the preparation of C^{14}.

* Present address: Mallinckrodt Institute of Radiology, Washington University, St. Louis, Mo.

1 Barker, H. A., and Haas, V., *Jour. Bact.*, **47**, 301–305 (1944).

2 Donker, H. J. L., Thesis, Delft (1926).

3 Schicktanz, S. T., Steeles, W. I., and Blaisdell, A. C., *Ind. Eng. Chem., Anal. Ed.*, **12**, 320–324 (1940).

4 Barker, H. A., and Kamen, M. D., these PROCEEDINGS, **31**, 219–225 (1945).

5 Wood, H. G., Brown, R. W., Werkman, C. H., and Stuckwisch, C. G., *Jour. Am. Chem. Soc.*, **66**, 1812–1818 (1944).

6 Buchanan, J. M., Hastings, A. B., and Nesbett, F. B., *Jour. Biol. Chem.*, **150**, 413–425 (1943).

7 Barker, H. A., Kamen, M. D., and Bornstein, B. T., these PROCEEDINGS, in press.

8 Barker, H. A., Ruben, S., and Beck, J. V., *Ibid.*, **26**, 477–482 (1940).

9 Wieringa, K. T., *Leeuwenhoek*, **6**, 251–262 (1939–1940).

POLLEN-TUBE GROWTH IN INTERGENERIC POLLINATIONS ON DATURA STRAMONIUM

BY CARMEN SANZ*

DEPARTMENT OF BOTANY, SMITH COLLEGE

Communicated October 15, 1945

During the past 30 or 40 years, research in pollen-tube growth has been carried out by numerous investigators. Pollen has been successfully germinated on artificial media and in a few cases on stigmas of unrelated plants (Yasuda[6] and Eigsti[5]). The rate of pollen-tube growth has been studied especially in relation to self-sterility. In many cases the pollen tubes grew slower in self-pollinations of self-incompatible varieties, so that pollen tubes did not reach the ovary before the styles decayed. In other cases pollen-tube development was found to be the same in both self and cross pollinations. The rate of pollen-tube growth in incompatible interspecific crosses has frequently been shown to be slower than in the compatible crosses. In some of the latter cases, an acceleration in the speed of the tubes was observed as they approached the ovary.

Studies of pollen-tube behavior following intergeneric pollinations have been carried out by several investigators. Yasuda[6] reports variation in the rate of pollen-tube growth in such crosses. He found faster tube growth in pollinations of *Petunia violacea* on pistils of *Solanum Gilo* than on those of *Solanum Melongena*, and also in pollinations of *Callistegia japonica* on *Cucurbita maxima* than on *Helianthus annuus*. Eigsti[5] reports normal growth of tubes of *Reseda* in styles of *Datura stramonium*.

The present research deals with pollen-tube behavior following intergeneric pollinations of *Datura stramonium*. Flowers of this species, Line 1, were used as females. The line is presumably homozygous, having originated from a duplication of the chromosomes of a haploid plant. In the following experiments, ten flowers were pollinated with pollen of each of the plants tested. All the experiments were made at a constant temperature of 26°C. Precautions in pollinations were observed, and preparation of microscopic slides by dissection, staining and mounting of stylar tissue of the pistils was made according to the techniques used by Buchholz.[1] By examination of these slides the extent of growth and behavior of the pollen tubes were determined. The number of normal pollen tubes and of abnormal ones (bursting or swollen) was observed and counted. An index for the speed of the normal-growing pollen tubes was calculated according to the methods used by Buchholz and co-workers.[2] On the basis of pollen germination and of the percentage of normal-growing pollen tubes, the species belonging to the Solanaceae and to other families were separated into the following classes showing: (*a*) no pollen germination; (*b*) most tubes

TABLE 1

CLASSIFICATION OF THE TESTED PLANTS ACCORDING TO THE GERMINATION AND THE PERCENTAGE OF NORMAL-GROWING POLLEN TUBES

FAMILIES	(a) NO GERMINATION	(b) MOST TUBES BURST	(c) ± HALF TUBES GROW	(d) MOST TUBES GROW
Solanaceae	*Nierembergia rivularis*	*Lycopersicum esculentum*	*Browallia speciosa*	*Brunfelsia calycina*
			Brunfelsia nitida	*Capsicum grossum*
			Lycium halimifolium	*Datura stramonium*
			Nicotiana glutinosa	*Nicotiana glauca*
			Physalis ixocarpa	*Nicotiana Langsdorffii*
			Petunia axillaris	*Nicotiana Tabacum*
			Solanum Melongena	*Salpiglossis sinuata*
Acanthaceae	*Ruellia Devosiana*			
Balsaminaceae	*Impatiens sultani*			
Campanulaceae	....	....	....	*Campanula carpatica*
Commelinaceae	*Tradescantia fuscata*			
	Tradescantia virginiana			
Compositae	*Catananche caerulea*			
	Rudbeckia hirta			
Convolvulaceae	*Convolvulus mauritanicus*			
Crassulaceae	*Sempervivum Moggridgei*			
Cucurbitaceae	*Cucurbita maxima*			
Gentianaceae	....	....	....	*Exacum affine*
Geraniaceae	*Geranium sessiliflorum*			
Gesneriaceae	*Achimenes longiflora*			
	Saintpaulia ionantha			
	Sinningia speciosa			
	Streptocarpus kewensis			

Hippocastanaceae	*Aesculus parviflora*			
Labiatae	*Salvia fulgens*	....	....	*Stachys discolor*
Leguminosae	*Swainsona galegifolia*			
Linaceae	*Linum alpinum*			
Lythraceae	*Lythrum salicaria*			
Malpighiaceae	*Thryallis glauca*			
Malvaceae	*Althaea rosea*	....	*Abutilon megapotamicum*	
	Hibiscus Rosa-sinensis			
Onagraceae	*Fuchsia speciosa*			
	Oenothera fruticosa			
Plumbaginaceae	*Plumbago Larpentae*			
Polemoniaceae	*Phlox paniculata*			
Portulacaceae	*Portulaca grandiflora*			
Primulacae	....	*Anagallis linifolia*		
Ranunculaceae	....	....	....	*Cimicifuga cordifolia*
				Ranunculus anemonoides
Rosaceae	*Potentilla fruticosa*			
	Geum grandiflorum			
Saxifragaceae	*Heuchera sanguinea*			
Scrophulariaceae	....	....	....	*Digitalis ambigua*
				Pentstemon Digitalis
Ternstroemiaceae	....	*Camellia japonica*		
Amaryllidaceae	*Alstroemeria aurantiaca*	....	*Narcissus Pseudo-Narcissus*	
Gramineae	*Zea mays*			
Iridaceae	....	*Crocus biflorus*	*Gladiolus Lemoinei*	*Freesia wisteria*
Liliaceae	*Lilium tigrinum*	....	*Hyacinthus myosotis*	*Tulipa Gesneriana*

bursting; (*c*) about an equal number of tubes growing and bursting, and (*d*) most tubes growing. The classes (*b*), (*c*) and (*d*) above are characterized by the following percentages of normal-growing pollen tubes: less than 45%, from 45 to 65%, and more than 65%, respectively.

Eighteen-Hour Test of Pollen-Tube Growth.—Styles of *Datura stramonium* were pollinated with pollen of 16 species of Solanaceae and 49 species of 31 other families. The growth and percentages of normal and bursting tubes were studied after 18 hours.

A classification of the tested plants into the four previously mentioned groups according to the pollen germination and to the percentages of normal-growing pollen tubes on pistils of *Datura stramonium* is given in table 1. It can be seen that the pollen of members of the Solanaceae more closely approached normal behavior. Most interest, however, should be attached to those cases of unrelated pollens which fall into the two top classes having 50% or more of normal-growing pollen tubes.

Tests for More Extended Periods of Time.—To determine whether or not normal pollen tubes observed in the 18-hour test would reach the ovary if given more time, some experiments were carried out in which the pollen tubes were allowed to grow for 66 hours before examination. The results of these tests are included in table 2. It can be seen that in two cases the earlier speed was approximately maintained. In most cases, however, the speed of growth declined considerably. With some of the species, tests were carried out for longer intervals of time until the styles began to rot at the base (*Petunia axillaris, Browallia speciosa, Freesia wisteria, Salpiglossis sinuata*). In all cases the pollen tubes were found to have stopped or to have practically stopped at some distance from the ovary.

Attempts to Induce the Pollen Tubes to Reach the Ovary.—*Tetraploid versus diploid styles:* In an attempt to discover a procedure to speed up the pollen-tube growth, tests were made with pollen of five plants on styles of a tetraploid *Datura stramonium*, Line 1, and the results were compared with those of the diploid styles of the same species, as shown in table 3. Most of the tubes burst when pollen of tomato (*Lycopersicum esculentum*) was used on diploid styles, but most of them grew when tetraploid styles were used. In all but one of the other cases some alleviation of the bursting was noted but to a much less degree. In all five cases growth was slightly increased in the tetraploid styles. However, the most striking effect was seen in the prevention of bursting in tomato.

Style splicing and style mutilation: Some methods to shorten the distance that the pollen tubes have to cover en route to the ovary were tried. The style splicing technique, described by Buchholz and co-workers,[3, 4] was used with some modifications. More promising results, however, were obtained with style mutilations. Pollinations were made inside vertical slits cut in the tops of the decapitated styles. When this technique was used

pollen tubes were induced to grow nearer to the ovary than in uncut styles. Pollen tubes of *Freesia wisteria*, of the Monocotyledoneae, were able to

TABLE 2

PERCENTAGE AND GROWTH OF NORMAL POLLEN TUBES OBSERVED AFTER 18 HOURS, AND OBSERVED AND CALCULATED AFTER 66 HOURS

FAMILIES	SPECIES	PERCENTAGE OF NORMAL TUBES	GROWTH IN 18 HOURS	GROWTH IN 66 HOURS	CALCULATED GROWTH IN 66 HOURS
Solanaceae	*Browallia speciosa*	59.4	7.1	8.0	25.5
	Brunfelsia calycina	88.8	13.7	18.8	49.3
	Brunfelsia nitida	46.3	19.7		
	Capsicum grossum	79.7	2.7		
	Datura stramonium	86.1	42.0		
	Lycium halimifolium	63.6	14.6		
	Lycopersicum esculentum	27.1	13.7		
	Nicotiana glauca	94.6	4.5	16.0	16.2
	Nicotiana glutinosa	63.1	9.0		
	Nicotiana Langsdorffii	90.9	10.7		
	Nicotiana Tabacum	87.1	17.7		
	Petunia axillaris	52.2	6.8	10.0	24.4
	Physalis ixocarpa	47.4	3.0	5.0	10.8
	Salpiglossis sinuata	88.2	18.6	20.0	66.9
	Solanum Melongena	58.3	7.1	12.0	25.5
Other Dicotyledoneae:					
Campanulaceae	*Campanula carpatica*	74.6	10.5		
Cucurbitaceae	*Cucurbita maxima*	6.8	1.8		
Gentianaceae	*Exacum affine*	70.9	5.5		
Labiateae	*Stachys discolor*	74.1	6.3		
Malvaceae	*Abutilon megapotamicum*	59.8	6.8		
Primulaceae	*Anagallis linifolia*	29.7	7.6		
Ranunculaceae	*Cimicifuga cordifolia*	90.0	2.4	27.7	27.8
	Ranunculus anemonoides	83.8	1.5		
Scrophulariaceae	*Digitalis ambigua*	84.6	4.2		
	Pentstemon Digitalis	90.7	12.3		
Terntroemiaceae	*Camellia japonica*	32.0	3.5		
Monocotyledoneae:					
Amaryllidaceae	*Narcissus Pseudo-Narcissus*	46.2	0.6		
Iridaceae	*Crocus biflorus*	25.0	0.6		
	Freesia wisteria	69.6	6.0	15.0	25.2
	Gladiolus Lemoinei	52.3	6.8		
Liliaceae	*Hyacinthus myosotis*	54.5	0.6		
	Tulipa Gesneriana	71.5	7.1		

reach the ovary, although they did not enter it. Pollen tubes of *Petunia axillaris*, *Salpiglossis sinuata* and *Browallia speciosa* grew closer to the ovary than on uncut styles examined when they began to rot at the base.

Summary.—In attempts to obtain intergeneric crosses, many barriers have been found which make these crosses unsuccessful. In this paper some of the barriers connected with pollen germination and with pollen-tube growth have been studied. All tests were made on stigmas and styles of *Datura stramonium*.

The pollens of members of the Solanaceae germinated more successfully than those of other families (93% against 43%). The fact that the pollen of species belonging to such distant groups as the Monocotyledoneae germinated on stigmas of *Datura stramonium* appears to indicate that pollen germination is not determined exclusively by taxonomic relationships.

On the average, the Solanaceae showed faster pollen-tube growth than members of other families (10.8 mm. against 5 mm.). However, in the latter group individual cases of relatively rapid growth of pollen tubes were found.

TABLE 3

FAMILIES AND SPECIES	PERCENTAGE OF NORMAL TUBES 2n	GROWTH 2n	PERCENTAGE OF NORMAL TUBES 4n	GROWTH 4n
Solanaceae				
Lycopersicum esculentum	**27.1**	**13.7**	**80.8**	**15.3**
Petunia axillaris	52.2	6.8	55.6	10.3
Physalis ixocarpa	47.4	3.0	75.6	3.8
Salpiglossis sinuata	88.2	18.6	84.7	25.6
Terntroemiaceae				
Camellia japonica	32.0	3.5	37.4	5.7

The use of tetraploid styles increased slightly the speed of the pollen tubes and decreased the percentage of bursting pollen tubes, especially in tomato, thereby increasing the chances for the tubes to reach the ovary.

Attempts to shorten the distance from the stigma to the ovary induced the pollen tubes to approach closer to the ovary. In one case the tubes arrived at the ovary, but in none did they enter it.

* Contributions from the Department of Botany, Smith College, New Series No. 18. Article taken from a thesis submitted to the faculty of Smith College in partial fulfilment of the requirements for the degree of Master of Arts in Botany under the direction of Dr. A. F. Blakeslee.

[1] Buchholz, J. T., "The Dissection, Staining and Mounting of Styles in the Study of the Pollen-Tube Distribution," *Stain Technology*, **6** (1), 13–24 (1931).

[2] Buchholz, J. T., and Blakeslee, A. F., "Pollen-Tube Growth at Various Temperatures," *Amer. Jour. Bot.*, **14** (7), 358–369 (1927).

[3] Buchholz, J. T., Doak, C. C., and Blakeslee, A. F., "Control of Gametophytic Selection in *Datura* through Shortening and Splicing of Styles," *Bull. Torrey Bot. Club*, **59** (3), 109–118 (1932).

[4] Buchholz, J. T., Doak, C. C., and Blakeslee, A. F., "Results Obtained from Splicing Styles," *Amer. Natur.*, **68**, 162–163 (1934).

[5] Eigsti, O. J., "Pollen-Tube Behavior in Self-Fertile, Self-Sterile and Interspecific Pollinated *Reseda*," *Amer. Natur.*, **71**, (736), 520–521 (1937).

[6] Yasuda, S., "On the Behavior of Pollen Tubes in the Production of Seedless Fruits Caused by Interspecific Pollination," *Japanese Jour. Genetics*, **8** (4), 239–244 (1933).

CYCLES IN MEASLES AND CHICKEN POX

BY EDWIN B. WILSON AND OLIVE M. LOMBARD

HARVARD SCHOOL OF PUBLIC HEALTH

Communicated September 17, 1945

Persons trained in mathematical physics or astronomy are apt to look for periods in oscillating phenomena and the methods of analysis (such as the periodogram) which have been developed by them for the determination of unknown periods are presumably well suited to eliminate efficiently the random fluctuations of phenomena which do really have an underlying period of unchanging length and phase. It is somewhat doubtful whether epidemiological or economic phenomena have any important periodic components except the seasonal or other variations impressed upon them by major astronomical periods, and it is presumably better that the more general term cycle be used for their ups and downs.[1]

In table 1 are given the reported cases of measles and of chicken pox in eleven cities for nineteen epidemic years.[2] For chicken pox the years are not the same for all cities. The epidemic year for these diseases is taken to run from September 1 to August 31. The number of turning points, as defined by Wallis and Moore,[3] is tabulated in the last row indexed *TP*. A downward or upward phase (W&M, p. 4) lies between two successive turning points and two successive phases form a cycle. The number of phases is thus one less than the number of turning points and the number of cycles is one-half of the number of phases. For measles the number of turning points averages 13.2, the number of phases 12.2 and the number of cycles 6.1. For chicken pox the number of turning points averages 12.9, the number of phases 11.9 and the number of cycles 6.0. The results are equivalent in view of the sampling errors.

The number of phases of 1, 2, 3, ... years duration can be enumerated and compared with the number that should be obtained for eleven series of nineteen items by chance (W&M, p. 10). It will be observed that both diseases have more phases than would occur by chance on the average. The number of phases can be regarded as normally distributed (W&M, p. 32) about 10.3 with variance[4] 3.06 and the sum of eleven series should be distributed about 113.7 with variance 33.61 or standard deviation 5.80, whereas the actual deviation is 20 for measles and 17 for chicken pox.

The result is therefore very significant; the ups and downs of measles and chicken pox in successive epidemic years are more frequent than in a chance series.

TABLE 1

REPORTED CASES OF MEASLES AND CHICKEN POX IN ELEVEN CITIES FOR NINETEEN EPIDEMIC YEARS, SEPTEMBER 1 TO AUGUST 31, BEGINNING ON SEPTEMBER 1 OF THE YEAR GIVEN

YEAR	CHIC.	CLEV.	DETR.	L. A.	MILW.	MONT.	N. Y.	PHIL.	PITT.	S. F.	TOR.*
					Measles						
1921	10,845	5,838	6,866	165	538	2281	41,478	3,492	3278	303	5,001
1922	15,878	6,516	4,004	3,638	10,941	2264	12,722	26,406	9220	1,954	2,834
1923	5,412	2,392	4,334	5,051	717	1019	35,311	3,655	855	3,853	1,454
1924	13,992	500	515	899	7,468	3783	4,571	7,703	7698	220	7,260
1925	5,324	15,865	16,520	440	4,242	1223	44,973	13,032	2987	2,707	2,191
1926	19,406	206	414	11,064	3,410	2300	1,724	949	2563	3,951	4,032
1927	1,202	2,589	15,321	672	191	4562	34,758	15,609	5973	817	2,360
1928	17,985	11,266	2,712	604	13,835	2138	3,100	1,450	1180	186	7,226
1929	1,131	321	15,153	6,650	437	2445	22,999	4,926	5365	7,558	660
1930	12,126	3,730	787	3,104	4,977	6450	27,072	15,705	2274	1,042	1,688
1931	9,749	15,097	12,836	456	14,497	4724	8,634	339	6204	3,775	10,208
1932	8,875	110	9,319	9,582	143	534	37,953	6,289	170	116	132
1933	9,379	3,693	2,004	1,142	2,500	4009	4,939	19,514	4388	4,426	78
1934	24,636	7,872	27,694	1,540	9,843	9366	27,782	1,319	9917	1,454	16,110
1935	541	2,346	737	9,210	321	3372	37,080	11,609	661	7,637	1,313
1936	3,995	5,582	1,220	677	371	6524	12,218	785	3256	202	3,707
1937	39,696	6,975	31,074	1,110	27,922	1817	34,624	15,740	7829	145	1,013
1938	572	202	741	9,377	158	9461	3,824	1,211	98	10,435	10,382
1939	2,169	163	4,047	604	3,284	1691	6,186	2,769	104	149	190
TP	14	11	15	11	12	12	15	14	14	13	14
					Chicken Pox						
1921	..	..	..	..	..	539	5,313	3916	..	..	..
1922	3915	..	..	..	..	577	8,469	2753	2015	..	..
1923	5713	3103	3130	3127	2936	794	7,959	5662	2806	1316	2098
1924	4054	3536	2400	2095	1902	596	7,720	3314	2345	1085	1777
1925	4794	1861	2276	1983	3907	1170	6,506	4634	1340	1640	2565
1926	4462	4000	3742	2197	3449	604	10,117	4651	2495	1511	2390
1927	4128	2149	1882	2569	2729	787	6,195	3307	1190	2566	2781
1928	4756	3887	3905	2523	4028	1565	9,764	4392	1875	894	2454
1929	4977	5101	3108	2306	5534	2802	7,939	3762	1771	1778	4233
1930	4934	5897	4244	2032	4789	2528	9,734	4993	2608	1658	3256
1931	4104	3623	2440	3950	2898	2147	7,763	3968	1809	2228	1950
1932	6953	7356	4769	2320	7648	3516	12,463	5937	3341	2201	4169
1933	6206	4665	3168	2754	6511	3667	11,847	4535	2600	2252	2620
1934	5727	5550	3977	3205	3364	4438	10,965	6474	2703	3336	4051
1935	6453	4615	4135	2025	9003	3193	8,180	4072	1939	1580	4548
1936	6356	4627	4611	3442	5552	4898	13,112	6771	3101	2994	4462
1937	6382	4394	3651	3194	3560	3303	10,852	6387	2037	1946	2450
1938	6010	4132	3323	2847	6002	3862	9,019	5606	2553	2612	3585
1939	7776	4947	3436	2176	5032	5068	12,069	6465	2394	1957	4629
1940	6090	4328	4141	2996	5834	..	..	..	2688	3516	2572
1941	..	5062	3096	4270	5807	..	..	..	..	2092	4141
TP	13	14	12	9	12	10	12	15	16	16	13

* The cities are Chicago, Cleveland, Detroit, Los Angeles, Milwaukee, Montreal, New York, Philadelphia, Pittsburgh, San Francisco, Toronto.

For a chance series there is a definite average duration of a phase (W&M, p. 10); with N = 19 it is 1.46. If we double this to get the average length

of a cycle we should have 2.92. The variance of the duration of phase is 0.51. The average duration for measles calculated[5] from table 2 is 1.28 and for chicken pox 1.31. This again shows that the oscillations are more frequent than they would be in a chance series. The differences $1.28 - 1.46 = -0.18$ and $1.31 - 1.46 = -0.15$ are both significant statistically. It should be especially noted that a chance series would have a mean duration of phase which would not be very different from the essentially equal values found for the mean duration in measles and in chicken pox; but our series involving eleven cities for nineteen years is long enough to show the significance of the difference between the observed mean and the theoretical mean. If we double the mean length of the phase to obtain a mean length of the cycle we find just over two and a half years for both measles and chicken pox.

It is characteristic of chicken pox and measles that the seasonal variation in reported cases is great, there being very few reported during the summer. The amount of measles varies so greatly between epidemic years that the epidemic nature of the disease is apparent despite the large seasonal fluctuation; but the variation of chicken pox from one epidemic year to another

TABLE 2

FREQUENCY OF PHASE DURATIONS

DURATION	1	2	3	ABOVE 3	TOTAL
Measles	98	35	1	0	134
Chicken Pox	99	24	8	0	131
Chance	73.3	30.2	8.1	2.0	113.7

is small enough so that one might be tempted to infer that the disease was essentially seasonal and otherwise non-epidemic. It, of course, is quite obvious that the intensity of the epidemic waves is far more pronounced in measles than in chicken pox; the reality of the epidemic character of chicken pox is, however, reasonably clear if one judges by the deviation of the mean duration of phase from that due to chance or from the excess of the number of phases compared with chance.

If table 2 for the frequency of phases of different durations for measles and chicken pox be treated as a contingency table and χ^2 be computed, we find $\chi^2 = 7.5$ for two degrees of freedom which would be formally significant —chiefly because of the relatively great differences in the numbers of phases of three years' duration. Thus there seems to be a slight difference between the two diseases in that chicken pox has more three-year phases than measles and possibly fewer two-year phases (the items might well be correlated).

Wallis and Moore (p. 39) state that their method is not very sensitive to a trend in the data; there is no doubt that there is trend in the reported cases from some cities but we have seen no evidence that the trend is suffi-

	OBSERVED			EXPECTED		
Measles	98	35	1	99.6	29.8	4.55
Chicken Pox	99	24	8	97.4	29.2	4.45

ciently marked to make its elimination important in the application of an insensitive test. If one proceeds to moving averages, thus reducing the amplitude of the variations, it is highly probable that the test might be less insensitive to some general underlying trend. We have, however, ventured to set up the tables for two-year and three-year moving averages and to tabulate the frequency of phases of different duration (table 3). It will

TABLE 3

FREQUENCY OF PHASE DURATIONS IN ONE-, TWO- AND THREE-YEAR MOVING AVERAGES

DURATION	1	2	3	ABOVE 3	TOTAL	MEAN
			Original Data (Single Years)			
Measles	98	35	1	0	134	1.28
Chicken Pox	99	24	8	0	131	1.31
Chance	73	30	8	2	114	1.46
			Two-Year Moving Average			
Measles	52	41	4	1	98	1.53
Chicken Pox	39	32	7	4	82	1.77
Chance	69	28	8	2	106	1.46
			Three-Year Moving Average			
Measles	79	15	9	2	105	1.39
Chicken Pox	80	7	8	5	100	1.42
Chance	64	26	7	2	99	1.46

be observed that for both moving averages in the case of measles the number of phases approaches that due to chance more closely than for the original data; the number of phases in the two-year moving average for chicken pox is notably short of the chance mark. The average duration of phase is also close to the chance value in all cases except the two-year moving average for chicken pox. In the moving averages there tends to be an excess of long phases as compared with the chance distribution or with that of the original data—possibly a reflection of trends in the data, though we have been unable to convince ourselves that it is. There seems to be no clear indication that the actual frequency distributions for the two-year or three-year moving averages are closer to the chance distributions than those for the single years.[6] Moreover, if the phenomenon is cyclic, rather than periodic, there is no obvious reason why the use of a moving average closer to the average length of the cycle, when the standard deviation of the length of a cycle is more than one-third of its mean, should be effective in bringing the frequency distribution close to those due to chance.

[1] Wilson, E. B., *Quarterly Jour. Economics*, **48**, 375–417 (1934); *Science*, **80**, 193–199 (1934).

[2] Dr. Mildred Wells, incident to a discussion of seasonal variation, *Am. Jour. Hyg.*, **40**, 279–317 (1944), has collected reported cases of some of the children's diseases by months in a large number of states and cities for twenty consecutive years and kindly sent us the figures. We have merely combined the monthly data into those for epidemic years.

[3] Wallis, W. A., and Moore, G. H., *A Significance Test for Time Series and Other Ordered Observations*, National Bureau of Economic Research, New York, 1941, 59 pages. We shall have to assume that this monograph is available to the reader, for we cannot take the space to reproduce the definitions and results we used; it will be cited as W&M.

[4] The variances actually observed between the eleven cities are 2.16 for measles and 5.09 for chicken pox, their average is 3.62; their ratio is 2.36 which for 10 degrees of freedom in each is not significant.

[5] Table 2 gives the total enumeration for the eleven cities. If the average duration is found for each city, we have as a grand average 1.29 for measles and 1.34 for chicken pox. (The average of the averages is not quite the same as the average for the total data because the number of phases differs.) The standard deviation of the eleven averages about their mean is for measles 0.14 (variance 0.019) and for chicken pox 0.21 (variance 0.044). The chance variance of 0.51 for the duration of phase would give a chance variance of 0.047 for the average duration of phase for 11 cities; this is near the variance found for chicken pox but about twice that found for measles. If we estimate the standard deviation of the means of eleven cities from the observed standard deviations we have 1.29 ± 0.042 and 1.34 ± 0.063. If we use table 2 we find variances of 0.22 for measles and 0.34 for chicken pox which for averages of about 130 elements would indicate 1.28 ± 0.040 and 1.31 ± 0.051. Thus the results of working with the eleven observed means and with the overall distribution are consistent. The differences $1.28 - 1.46 = -0.18 \pm 0.040$ and $1.31 - 1.46 = -0.15 \pm 0.051$ are significant; if we use the values obtained by working with the eleven cities namely, $1.29 - 1.46 = -0.17 \pm 0.042$ and $1.34 - 1.46 = -0.12 \pm 0.063$ the result for chicken pox barely escapes significance.

[6] We offer this as a statistical judgment only. W&M discuss significance tests at length with reference to χ^2. If the values of χ^2 are computed from the four classifications of phase, viz., one, two, three and more than three years of duration, despite the small theoretical number 2 in the chance series, we find for measles 18, 13, 9 for single years, two-year, three-year moving averages and for chicken pox 12, 16, 23; measles seems to be improving, chicken pox worsening.

PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 31 December 15, 1945 Number 12

THE SYNTHESIS OF BUTYRIC AND CAPROIC ACIDS FROM ETHANOL AND ACETIC ACID BY CLOSTRIDIUM KLUYVERI

BY H. A. BARKER, M. D. KAMEN* AND B. T. BORNSTEIN

DIVISION OF PLANT NUTRITION, UNIVERSITY OF CALIFORNIA

Communicated October 29, 1945

Unpublished investigations in this laboratory have shown that *Clostridium kluyveri* can metabolize acetic acid and ethanol under anaerobic conditions producing butyric and caproic acids in accordance with equations (1) and (2).

$$CH_3CH_2OH + CH_3COOH = CH_3CH_2CH_2COOH + H_2O \quad (1)$$

$$2CH_3CH_2OH + CH_3COOH = CH_3(CH_2)_4COOH + 2H_2O \quad (2)$$

The relative yields of the two acids are determined by the relative amounts of acetic acid and ethanol available. If acetic acid is present in excess, a considerable amount of butyric acid is formed, while if ethanol is in excess, caproic acid is the main product. These relations suggest that butyric acid may be an intermediate in the synthesis of caproic acid from acetic acid. This is supported by the observation that ethanol and butyric acid can be converted to caproic acid according to equation (3).

$$CH_3CH_2OH + CH_3CH_2CH_2COOH = CH_3(CH_2)_4COOH + H_2O \quad (3)$$

In the present investigation we have obtained positive proof for the conversion of acetic acid to caproic acid via butyric acid by studying the action of *Cl. kluyveri* on media containing synthetic fatty acids labeled with the long-lived radioactive carbon isotope C^{14}. In addition we have obtained various types of evidence which help to elucidate the mechanism of fatty acid synthesis from C_2 molecules.

Experimental Methods.—Carboxyl-labeled acetic, butyric and caproic acids were prepared from C^*O_2 and methyl iodide, propyl bromide and amyl bromide, respectively, by the Grignard synthesis.[1]

The general technique of estimating the C^{14} content of carbon dioxide and fatty acids has already been described elsewhere.[2] The barium salts

of the fatty acids dried at 100°C. were used in the radio-assay. The individual acids were separated from mixtures by the distillation procedure of Schicktanz, *et al.*[3] The purity of each acid was established by Duclaux distillation. The methods used in locating C^{14} in the individual carbon atoms of acetic and butyric acids have been described previously.[2, 4, 5] The occurrence of C^{14} in the carboxyl group of caproic acid was determined by decarboxylating the barium salt. Control experiments with synthetic caproic acid showed that this method yields 52 per cent of the carboxyl carbon as barium carbonate.

The bacteria were grown in media of the following composition in g. per 100 ml.: ethanol 0.3–0.8, sodium salt of labeled acetic or butyric acid 0.3–0.6, yeast autolysate 0.3, pH 7.2, phosphate buffer $M/40$, $(NH_4)_2SO_4$ 0.1, $MgSO_4 \cdot 7H_2O$ 0.01, $FeSO_4.7H_2O$ 0.0001, sodium thioglycollate 0.05,

TABLE 1

THE FERMENTATION OF CH_3CH_2OH AND CH_3C^*OOH BY *Cl. kluyveri*

Experiment 1

COMPOUND	MM	CTS./MIN./MM	TOTAL CTS./MIN.
Before growth			
Ethanol	ca. 1.1	....	....
Acetic acid	0.830	18,900 ± 300	15,700
Carbon dioxide	ca. 0.2	....	....
After growth			
Acetic acid	0.395	7330 ± 150	2895
Butyric acid	0.196	18,200 ± 300	3570
Caproic acid	0.257	24,500 ± 500	6300
Carbon dioxide	0.185	26 ± 50	5
Ethanol	0.217	260 ± 50	74
Non-volatile cpds.	...	....	193
Percentage recovery of C^{14}			83

traces of salts of Ca, Mn and Mo, distilled water. In some experiments sodium carbonate (0.025 g. per 100 ml.) was added as a sterile solution after autoclaving the medium. Oxygen was removed by means of either a pyrogallol-potassium carbonate or an Oxsorbent seal. The cultures were incubated at 32–35° until growth ceased before being analyzed.

Results.—When *Cl. kluyveri* was allowed to grow in a medium containing approximately equivalent amounts of ethanol and synthetic carboxyl-labeled sodium acetate, radioactive butyric and caproic acids were formed (table 1). The isotope was proved to be present in these acids, rather than in some associated compound, by establishing the constancy of the specific activities of the barium salts prepared from different fractions of a Duclaux distillation (table 2). A little C^{14} was also found in ethanol and unidentified non-volatile compounds, but no significant amount was present in carbon dioxide. It should be noted that there is no net production of

carbon dioxide in this fermentation. The recovery of C^{14} in all forms after growth was about 83 per cent of that initially added as acetic acid.

A particularly significant fact revealed by this experiment is the decrease in the molar activity of acetic acid during the fermentation. At the beginning of the experiment the activity expressed in cts./min./mM was 18.9×10^3 while at the end the value was 7.3×10^3. The simplest and most likely explanation for this reduction in activity is the oxidation of ethanol to inactive acetic acid or some closely related compound in equilibrium with it. This would cause a dilution of the labeled acetic acid.

TABLE 2

Specific Activities of Barium Salts of Butyric and Caproic Acids Prepared from Duclaux Distillations

Experiment 1 (total volume = 110 ml.)

FRACTION	CTS./MIN./MG. OF Ba SALT
Butyric acid	
0–20 ml.	47.5
20–40 ml.	47.6
40–110 ml.	46.4
Caproic acid	
0–20 ml.	63.7
20–110 ml.	62.2

Another possible explanation for the decrease in activity of acetic acid is that it is in chemical and isotopic equilibrium with the higher fatty acids formed from it and ethanol. This was tested by two experiments in one of which butyric acid labeled in the carboxyl group was added at the beginning of the fermentation; in the second experiment butyric acid labeled on all four carbon atoms was used. In the experiment with carboxyl-labeled butyric acid (table 4), no C^{14} could be detected in the final acetic acid, thus indicating that no exchange occurred between the carboxyl groups of butyric acid and acetic acid. In the second experiment with completely labeled butyric acid, a little C^{14} was recovered in the final acetic acid but it was less than 9 per cent of the amount to be expected if butyric and acetic acids were in isotopic equilibrium, and even this small effect may have been due to a slight contamination of the original butyric acid by labeled acetic acid. An exchange of C^{14} between acetic acid and the higher fatty acids can therefore be excluded as an explanation for the observed decrease in activity of acetic acid.

A third possible explanation for this effect is an exchange of C^{14} between ethanol and acetic acid as a result of reactions involving acetaldehyde or a similar compound as an intermediate. If such a reaction occurs, the residual ethanol should have the same molar activity as the acetic acid. Actually, the activity of the alcohol is of a lower order of magnitude

(table 1). Therefore an exchange of C^{14} between acetic acid and ethanol is not consistent with the observed results.

Whatever the actual mechanism of the dilution of the active acetic acid, the magnitude of the effect is about what would be expected if all the ethanol is oxidized to acetic acid while the latter is being converted to higher fatty acids. To calculate the expected decrease of acetic acid activity on this basis, let us assume that (1) the fermentation proceeds in two successive steps represented by equations (1) and (3), and that during the fermentation all the ethanol is first converted to acetic acid or some compound in isotopic equilibrium with it. In step 1, the formation of butyric acid from ethanol and acetic acid, let x = the amount of C^{14} present in acetic acid at any time, x_0 = the initial C^{14}, x_1 = the C^{14} at the end of step 1, V = the amount of acetic acid present at any time, V_0 = the initial acetic acid, and V_f = the final acetic acid. Now during step 1, when a small amount (ΔV) of alcohol is oxidized to inactive acetic acid, the removal of active acetic acid by conversion to butyric acid will be equal to $2\Delta V$. The loss of C^{14} from acetic acid is therefore

$$-\Delta x = -2\Delta V/V - \Delta V\cdot x. \qquad (4)$$

In the limit as ΔV approaches zero

$$dx/x = 2dV/V. \qquad (4a)$$

Integrating between the limits of V_0 and V_f, and x_0 and x_1 and changing to $\log_{10}$ we get

$$\log x_1/x_0 = 2 \log V_f/V_0 \qquad (5)$$

or

$$x_1 = x_0(V_f/V_0)^2. \qquad (5a)$$

In step 2 (equation 3), the conversion of alcohol and butyric acid to caproic acid, there is no change in the total quantity of acetic acid; V_f is therefore a constant. C^{14} will nevertheless be lost from acetic acid if, as is assumed, it is an intermediate in the utilization of ethanol. When a small amount of alcohol (ΔA) is used, the loss of C^{14} from acetic acid is

$$-\Delta x = \Delta A/(V_f + \Delta A)\cdot x. \qquad (6)$$

In the limit

$$-dx/x = dA/V_f. \qquad (6a)$$

Integrating between the limits x_1 and x_f, and 0 and A_f, where A_f is the total amount of alcohol used in step 2

$$\ln x_1/x_f = 2.3 \log x_1/x_f = A_f/V_f \qquad (7)$$

or

$$\log x_f = \log x_1 - A_f/2.3V_f \qquad (7a)$$

When this method of calculation is applied to the data of table 1 where x_0 = 15,600 cts./min., V_0 = 0.830 mM, V_f = 0.395 mM and A_f = 0.257 mM, it is found that x_1 = 3540 cts./min., and x_f = 1840 cts./min. The calculated molar activity of acetic acid at the end of the fermentation is therefore 1840 cts./min./0.395 mM = 4660 cts./min./mM. This value is of the same order of magnitude though somewhat smaller than the observed molar activity of 7330 cts./min./mM. The discrepancy may be due to the fact that butyric and caproic acids are formed simultaneously rather than successively during part of the fermentation; this would raise the calculated activity of the final acetic acid and bring it into closer agreement with the observed value. The general conclusion to be drawn from a comparison of the calculated and observed decrease in the molar activity of acetic acid is that the results support the view that alcohol is oxidized to acetic acid or a closely related compound which is then condensed to butyric and caproic acids.

The following information was obtained concerning the distribution of C^{14} in the fatty acids at the end of the fermentation.

TABLE 3

DISTRIBUTION OF C^{14} IN BUTYRIC ACID

Experiment 2

CARBON ATOM	CTS./MIN./MM	PER CENT OF TOTAL RECOVERED
Carboxyl	1680 ± 50	54 ± 2
Alpha and gamma (average)	28 ± 50	1 ± 2
Beta	1395 ± 30	45 ± 2

The final like the initial acetic acid contained C^{14} only in the carboxyl group. The actual results were 7400 ± 200 cts./min./mM of carboxyl carbon and 0 ± 30 cts./min./mM of methyl carbon.

Data on the distribution of C^{14} in butyric acid are presented in table 3. The butyric acid was obtained from a second experiment identical with experiment 1. The data show that C^{14} is present in the carboxyl and beta positions but not in the alpha and gamma positions. The quantity of C^{14} in the carboxyl and beta carbon atoms is almost equal, though there appears to be a small but significantly greater amount in the carboxyl carbon. The smaller C^{14} content of the beta carbon may result from the fact that the beta (and gamma) carbon atoms are derived from an oxidation product of ethanol which is not in complete isotopic equilibrium with the preformed acetic acid.

No method was available for determining the complete distribution of C^{14} in caproic acid. The presence and amount of C^{14} in the carboxyl group could, however, be determined by decarboxylation of the barium salt. We found 4150 cts./min./mM in the barium carbonate obtained by the

decarboxylation of barium caproate containing 11,600 cts./min./mM. Therefore, 35.7 per cent or slightly more than one-third of the C^{14} was present in the carboxyl group. The data are consistent with the view that the C^{14} is equally distributed among three of the six carbon atoms (probably those in the carboxyl, beta and delta positions), but this is not certain.

The results of experiment 1 (table 1), considered in conjunction with other information now available, lead us to postulate that the following reactions may be involved in the formation of butyric and caproic acids from acetic acid and ethanol.

$$CH_3CH_2OH + K_2HPO_4 \rightarrow CH_3COOPO_3K_2 + 4H \qquad (8)$$

$$CH_3COOPO_3K_2 + ADP \rightleftharpoons CH_3COOH + ATP \qquad (9)$$

In the presence of synthetic CH_3C^*OOH, the forward reaction (9) would cause a dilution of the active acetic acid and the reverse reaction would introduce C^{14} into the carboxyl group of acetyl phosphate. Although there is at present no direct evidence to indicate the occurrence of acetyl phosphate in this fermentation, the compound is almost certainly formed by *Cl. butylicum*,[6] an organism which also produces butyric acid (and butyl alcohol) from acetic acid.[5] Moreover the chemical properties of acetyl phosphate are consistent with its being an intermediate in fatty acid synthesis.[7, 8] The labeled acetyl phosphate and the labeled acetic acid are assumed to condense to yield a product that is reduced to butyric acid. This reaction is not reversible.

$$CH_3C^*OOPO_3K_2 + CH_3C^*OOH + 4H \rightarrow CH_3C^*H_2CH_2C^*OOH + K_2HPO_4 + H_2O. \qquad (10)$$

A similar type of condensation and reduction involving butyric acid and acetyl phosphate could lead to the formation of caproic acid labeled in the carboxyl, beta and delta positions.

$$CH_3C^*OOPO_3K_2 + CH_3C^*H_2CH_2C^*OOH + 4H \rightarrow CH_3C^*H_2CH_2C^*H_2CH_2C^*OOH + K_2HPO_4 + H_2O. \qquad (11)$$

Reaction (11) represents only one of two possible ways in which caproic acid could be formed. Instead of there being a direct addition of the acetyl group to the terminal methyl group of butyric acid as is indicated above, the high energy phosphate group could be first transferred to butyric acid to give butyryl phosphate which could then condense with the methyl group of acetic acid.

$$CH_3C^*H_2CH_2C^*OOH + CH_3C^*OOPO_3K_2 \rightleftharpoons CH_3C^*H_2CH_2C^*OOPO_3K_2 + CH_3C^*OOH. \qquad (12)$$

$$CH_3C^*H_2CH_2C^*OOPO_3K_2 + CH_3C^*OOH + 4H \rightarrow CH_3C^*H_2CH_2C^*H_2CH_2C^*OOH + K_2HPO_4 + H_2O. \qquad (13)$$

Evidence has been presented by Koepsell, *et al.*,[6] that a transphosphorylation between acetylphosphate and butyric acid occurs under the influence of an enzyme preparation obtained from *Cl. butylicum*. With either mechanism the isotope distribution of the resulting caproic acid would be the same under the conditions of experiment 1.

It is possible, however, to distinguish between these two types of condensation by carrying out a fermentation of ordinary ethanol and synthetic butyric acid labeled in the carboxyl group. The position of the C^{14} in the resulting caproic acid must depend on the way in which the condensation occurs. This is illustrated in equations (14) and (15) where the inactive acetyl phosphate and acetic acid are assumed to originate from the oxida-

$$CH_3COOPO_3K_2 + CH_3CH_2CH_2C^*OOH + 4H \rightarrow CH_3CH_2CH_2CH_2CH_2C^*OOH + K_2HPO_4 + H_2O. \quad (14)$$

$$CH_3(CH_2)_2C^*OOPO_3K_2 + CH_3COOH + 4H \rightarrow CH_3(CH_2)_2C^*H_2CH_2COOH + K_2HPO_4 + H_2O. \quad (15)$$

tion of ethanol. It can be seen that a condensation of acetyl phosphate and labeled butyric acid would yield carboxyl-labeled caproic acid, while a condensation of labeled butyryl phosphate and ordinary acetic acid would yield caproic acid labeled in the beta position.

TABLE 4

THE FERMENTATION OF CH_3CH_2OH AND SYNTHETIC $CH_3CH_2CH_2C^*OOH$ BY *Cl. kluyveri*

Experiment 3

COMPOUND	MM/10 ML.	CTS./MIN./MM	TOTAL CTS./MIN.
Before growth			
Acetic acid	0.100	0	0
Butyric acid	0.384	17,200 ± 500	6610
Ethanol	ca. 1.7	0	0
After growth			
Acetic acid	0.101	0 ± 100	0
Butyric acid	0.05	18,000 ± 10,000	955
Caproic acid	0.427	12,600 ± 500	5400
Ethanol	0.306	250 ± 50	75

The results of an experiment designed to distinguish between these possible condensation mechanisms are given in table 4. Alcohol was added in excess to favor a high conversion of butyric acid to caproic acid.

It can be seen in table 4 that most (81.7 per cent) of the C^{14} originally present in butyric acid was converted into caproic acid. The molar activity of the caproic acid (12,600 ± 500 cts./min./mM) is somewhat less than that of the original butyric acid (17,200 ± 500 cts./min./mM) due to the fact that part of the caproic acid was completely built up from in-

active ethanol and acetic acid. This is evident from the fact that the quantity of caproic acid formed (0.427 mM) was greater than the butyric acid originally added (0.384 mM). When this dilution is taken into account, the calculated molar activity of the final caproic acid (17,200 cts./min./mM $\times$ 0.384 mM/0.427 = 13,400 cts./min./mM) agrees fairly well with its observed molar activity (12,600 cts./min./mM).

The caproic acid formed in this experiment contained little or no C^{14} in the carboxyl group. The barium carbonate obtained by decarboxylation from two samples of barium caproate showing activities of 845 $\pm$ 20 and 657 $\pm$ 15 cts./min., were found to give only 6 $\pm$ 3 and 5 $\pm$ 3 cts./min., respectively. It will be recalled that synthetic carboxyl-labeled barium caproate yields 52 per cent of its C^{14} as barium carbonate.

The absence of C^{14} from the carboxyl group of caproic acid proves that a condensation of the type shown in equations (11) and (14) does not occur. The results are, however, consistent with the mechanism shown in equations (13) and (15), involving a condensation of butyryl phosphate with the methyl group of acetic acid. The evidence in favor of this type of condensation would be stronger if the C^{14} had been proved to be present only in the beta position. This has not yet been accomplished, but further work on this problem is in progress.

There are a few other points concerning experiment 3 (table 4) that should be mentioned.

The complete absence of activity in acetic acid definitely proves that the conversion of acetic to butyric acid is not appreciably reversible at least so far as the carboxyl carbon atom is concerned. Evidence has already been presented above against any considerable exchange of carbon between acetic acid and the alpha, beta or gamma carbon atoms of butyric.

A small but definite amount of activity was found in the neutral volatile compounds, reported as ethanol. It is possible that this activity is actually present in ethanol, though this would be difficult to reconcile with the fact that there is no activity in acetic acid which seems to be an intermediate between butyric acid and ethanol. Another possibility is that the activity attributed to ethanol is actually present in an associated compound such as butyl or hexyl alcohol. Since the molar activities of the C_4 and C_6 compounds are relatively very high, a small (1–2 per cent) contamination of the ethanol would account for the observed activity. With the small amount of alcohol and the low activity involved in the present experiment, the contamination theory could not be proved or disproved. However, it may be mentioned in this connection that the characteristic odor of *Cl. kluyveri* cultures suggests the presence of a higher alcohol or ester.

Summary.—1. When *Cl. kluyveri* is grown anaerobically in a medium containing ordinary ethanol and synthetic acetic acid labeled in the carboxyl group with C^{14}, labeled butyric and caproic acids are formed.

2. The butyric acid so formed has C^{14} almost equally distributed between the carboxyl and beta positions. The alpha and gamma positions are inactive.

3. The caproic acid has one-third of its C^{14} in the carboxyl group; probably the beta and delta positions are also labeled, though this has not been proved.

4. No active carbon dioxide is formed from carboxyl-labeled acetic acid. This indicates that carbon dioxide is not an intermediate in these reactions.

5. The C^{14} content of the residual acetic acid is much lower than that of the initial acetic acid. This evidently results from the oxidation of ethanol to acetic acid or a related compound in isotopic equilibrium with it.

6. When *Cl. kluyveri* is grown with ordinary ethanol and synthetic carboxyl-labeled butyric acid, C^{14} is found in caproic acid but not in acetic acid.

7. The active caproic acid so formed contains almost no activity in its carboxyl group.

8. The above facts are consistent with the view that the formation of butyric acid involves a condensation between acetic acid or a reactive derivative thereof, such as acetylphosphate, formed by the anaerobic oxidation of ethanol, and a second molecule of acetic acid. The condensation product is then reduced to butyric acid. Caproic acid formation involves a condensation of the carboxyl group of butyric acid or some related C_4 compound, like butyrylphosphate, with the methyl group of acetic acid.

* Present address: Mallinckrodt Institute of Radiology, Washington University, St. Louis, Mo.

[1] Buchanan, J. M., Hastings, A. B., and Nesbett, F. B., *Jour. Biol. Chem.*, **150**, 413–425 (1943).

[2] Barker, H. A., and Kamen, M. D., these PROCEEDINGS, **31**, 219–225 (1945).

[3] Schicktanz, S. T., Steeles, W. I., and Blaisdell, A. C., *Ind. Eng. Chem., Anal. Ed.*, **12**, 320–324 (1940).

[4] Barker, H. A., Kamen, M. D., and Haas, V., these PROCEEDINGS, **31**, 355–360 (1945).

[5] Wood, H. G., Brown, R. W., Werkman, C. H., and Stuckwisch, C. G., *Jour. Am. Chem. Soc.*, **66**, 1812–1818 (1944).

[6] Koepsell, H. J., Johnson, M. J., and Meek, J. S., *Jour. Biol. Chem.*, **154**, 535–547 (1944).

[7] Lipmann, F., *Ibid.*, **153**, 571–582 (1944).

[8] Lipmann, F., *Advances in Enzymology*, Interscience Publishers, Vol. 1, 99–162 (1941).

THE DIFFERENTIAL EQUATION OF THE DISTRIBUTION OF GENE FREQUENCIES

BY SEWALL WRIGHT

DEPARTMENT OF ZOÖLOGY, THE UNIVERSITY OF CHICAGO

Communicated October 15, 1945

The first attempt to determine the mathematical form of the distribution of gene frequencies in populations was based on the setting up of differential equations for certain special cases (Fisher, 1922,[1] 1930[2]). A correction and extension of these results came from expression of the conditions in an integral equation (Wright, 1929,[3] 1931[4]). A general solution has since been obtained for fully stationary distributions by a third method (Wright, 1937,[5] 1938[6]). The case of uniform flux has been treated less generally (Wright, 1938,[6] 1942[7]). Dr. A. Kolmogorov[8] has recently been kind enough to send me a reprint of an important paper on this subject which was published in 1935 but which had not previously come to my attention. While the application is restricted to a particular stationary distribution, the method of approach points to a more systematic formulation than before.

The situation discussed by Kolmogorov is that of a large population, consisting of many subgroups of size n, each of which receives a certain number (k) of immigrants from the general population but otherwise breeds within itself. The average rate of change of the gene frequency of subgroups, in which p is the frequency of a given gene, is represented by $A = \Sigma(\Delta p) = (k/n)(\bar{p} - p)$ where $\bar{p}$ is the mean value of p in the whole population. The variance of p, due to accidents of sampling in one generation, is represented by $B = \Sigma(\Delta p)^2 = pq/2n$, $q = 1 - p$. It is stated, without demonstration, that the distribution $u(p)$ of gene frequencies among subgroups after a stationary state has been reached, answers to the differential equation

$$\frac{1}{2}\frac{\partial^2}{\partial p^2}(Bu) - \frac{\partial}{\partial p}(Au) = 0. \qquad (1)$$

The pertinent solution is given as

$$u(p) = p^{4k\bar{p}-1}q^{4k\bar{q}-1}/B(4k\bar{p}, 4k\bar{q}). \qquad (2)$$

The effect of selection in this situation is discussed briefly without, however, modifying $u(p)$ by introduction of the selection term, $\alpha p^2 q$, into A.

It is noted that the same formula (2) had previously been derived by the present author[4] by a different method. Equation (1) has, however, broader implications if valid for the general case $A = \Sigma(\Delta p)$, $B = \Sigma(\Delta p)^2$ and not merely for the particular case $A = (k/n)(\bar{p} - p)$, $B = pq/2n$.

The Immediate Factors of Evolutionary Change.—The immediate factors that tend to cause systematic changes (Δq) in gene frequency (q) may be listed exhaustively[4, 5] as (a) mutation pressure, $\Delta q = v(1 - q) - uq$, where v and u are the rates of mutation to and from the gene in question, (b) immigration pressure, $\Delta q = m(q_t - q)$ (Kolmogorov's A) where m is the proportion of replacement by immigrants and q_t is the gene frequency in these, and (c) selection pressure, which may take widely diverse forms but in the important case of constant relative selective values (W) for each multiple factor genotype in a random breeding population takes the form $\Delta q = q(1 - q)\dfrac{\partial \bar{W}}{\partial q}/r\bar{W}$, where r is 1 in haploids, 2 in diploids, the usual case, 1.5 for sex linked genes (if equal numbers of males and females), 4 in tetraploids, etc.[7, 9] In addition to these systematic pressures are (d) the random variations, δq, due to accidents of sampling, the variance of which is $\sigma^2_{\delta q} = q(1 - q)/rN$ in a population of effective size N (Kolmogorov's B). The diploid case ($r = 2$) will be assumed in what follows.

The Stationary Distribution of Gene Frequencies.—Systematic pressure toward the gene frequency, at which $\Delta q = 0$ and the cumulative effects of accidents of sampling determine a probability curve $\varphi(q)$ describing the frequencies which would be exhibited in the long run by the value of q for a particular gene in a population subject to constant conditions. This distribution may also be interpreted as that exhibited at one time by the values of q in a group of populations that are all subject to the same conditions (as in the case of Kolmogorov's $u(p)$). The deviations from the binomial square formula for genotypic frequencies in the total population, depend on the variance of $\varphi(q)$ under this interpretation.[4, 8, 10] In other cases $\varphi(q)$ may be used as the distribution at any time within either a class of non-allelic genes or an extensive series of multiple alleles,[11] all subject to the same conditions.

That equation (1) is, in fact, completely general for the stationary form of distribution may be shown by a slight modification of a method[5, 6] that has been used for derivation of $\varphi(q)$.

The conditions for stability of the distribution (including the terminal classes $q = 0$, $q = 1$) may be represented by two equations expressing the persistence of the mean and variance, respectively

$$\int_0^1 (q + \delta q + \Delta q)\varphi(q)dq = \int_0^1 q\varphi(q)dq. \tag{3}$$

$$\int_0^1 (q - \bar{q} + \delta q + \Delta q)^2\varphi(q)dq = \int_0^1 (q - \bar{q})^2\varphi(q)dq. \tag{4}$$

Noting that the mean value of δq is zero, and that δq is not correlated with q or Δq, these equations reduce to the following if the term in $(\Delta q)^2$ in (4) may be ignored. It may be noted in this connection that this term is negligible if Δq is of the same order as $\sigma^2_{\delta q}$ or less, while if of higher order,

systematic pressure dominates the results so completely that the distribution formula itself becomes unimportant.

$$\int_0^1 \Delta q \varphi(q) dq = 0. \tag{3a}$$

$$2 \int_0^1 (q - \bar{q}) \Delta q \varphi(q) dq + \int_0^1 \sigma_{\delta q}^2 \varphi(q) dq = 0. \tag{4a}$$

Putting $\Delta q \varphi(q) dq = d\chi(q)$ these conditions become

$$\chi(1) - \chi(0) = 0. \tag{3b}$$

$$2 \int_0^1 \chi(q) dq - 2[\chi(1) + \bar{q}(\chi(1) - \chi(0))] - \int_0^1 \sigma_{\delta q}^2 \varphi(q) dq = 0. \tag{4b}$$

Substituting (3*b*) in (4*b*) the latter becomes

$$\int_0^1 [2\chi(q) - 2\chi(1) - \sigma_{\delta q}^2 \varphi(q)] dq = 0. \tag{4c}$$

A solution is obtained by removing the integral sign since the resulting equation not only satisfies (4*c*) but also (3*b*) (noting that $\sigma_{\delta q}^2 = 0$ if $q = 0$ or if $q = 1$, there being no sampling variance without alternatives in the sample).

$$\chi(q) - \chi(1) = {}^1/_2 \sigma_{\delta q}^2 \varphi(q). \tag{5}$$

This can be solved for $\varphi(q)$ by differentiating the logarithm of the left-hand number and making the appropriate substitutions.[6]

$$\varphi(q) = (C/\sigma_{\delta q}^2) e^{2\int (\Delta q/\sigma_{\delta q}^2) dq} \tag{6}$$

where C is a constant such that $\int_0^1 \varphi(q) dq = 1$.

Since q increases by steps of $1/2N$ in a population of size N, the frequency of a given value of q is $f(q) = \varphi(q)/2N$. From a study[4] of simple cases ($N = 2$ or 3) in which the frequencies in the stationary state can be determined algebraically and from a more elaborate investigation by R. A. Fisher[2] of the subterminal region in certain cases, it appears that the frequencies are given with considerable accuracy by the formula except for the terminal classes, $q = 0$, $q = 1$. Consideration of the exchanges which occur between the terminal and neighboring classes leads[4] to the following approximate estimate for the terminal class, $q = 0$. That for $q = 1$ is analogous.

$$f(0) = f(1/2N)/4N[mq_t + v]. \tag{7}$$

The differential equation for the completely stationary case is given by differentiation of (5). It comes under equation (1).

$$\frac{1}{2} \frac{d}{dq} (\sigma_{\delta q}^2 \varphi(q)) - \Delta q \varphi(q) = 0. \tag{8}$$

Since $\Delta q \varphi(q)$ is the proportion of the distribution which tends to be carried past a specified value of q by the systematic pressure Δq, the other

term must represent the net proportion which tends to be carried in the opposite direction by accidents of sampling in each generation.

The Case of Steady Flux.—There may be a practically stationary state of the proportions in all intermediate values of q in spite of steadily increasing frequency of one terminal class at the expense of the other, provided that the proportion lost by the donor terminal class is negligible. This cannot be the case if either mutation rate or immigration rate is appreciable, but may hold in the presence of strong selection pressure since selection pressure is nil in populations in which $q = 0$ or $q = 1$.

The differential equation for the case of steady flux must differ from (8) by a constant term (D), the net proportion of the total (excluding the recipient class) that is carried past each value of q in each generation.

$$\frac{1}{2}\frac{d}{dq}(\sigma^2_{\delta q}\varphi(q)) - \Delta q\varphi(q) + D = 0. \qquad (9)$$

This is the general form given by one integration of (1) which is therefore the general differential equation for a steady state of the intermediate classes. It may be reduced to a linear equation of the first order by making the substitution, $y = \sigma^2_{\delta q}\varphi(q)$.

$$\frac{dy}{dq} - 2\left(\frac{\Delta q}{\sigma^2_{q}}\right)y + 2D = 0. \qquad (10)$$

The solution for $\varphi(q)$ is as follows:

$$\varphi(q) = [e^{2\int(\Delta q/\sigma^2_{\delta q})dq}/\sigma^2_{\delta q}]\,[C - 2D\int e^{-2\int(\Delta q/\sigma^2_{\delta q})dq}dq]. \qquad (11)$$

The simplest special case is that in which Δq may be treated as zero (although there could be no flux if it were absolutely zero).

$$f(q) = \frac{C}{q(1-q)} - \frac{2D}{1-q}. \qquad (12)$$

The case under (12) that is most important genetically is that of irreversible mutation at a rate so low that the donor class ($q = 0$, or $q = 1$) is not appreciably depleted. According to direction of mutation,

$$f(q) = 2v/q, \text{ or } f(q) = 2v/(1-q). \qquad (13)$$

The ratio of the subterminal classes ($1/2N$ in this case) gives the probability that a single neutral mutation may reach fixation instead of elimination.

Returning to (12) the case in which $D = 0$ yields the corresponding simplest solution for a completely stationary state

$$\varphi(q) = 1/[2(0.577 + \log 2N)q(1-q)] \text{ (terminal classes excluded)}. \qquad (14)$$

The case in which there are constant relative selection coefficients for all genotypes ($\Delta q = q(1 - q)\frac{\partial \overline{W}}{\partial q}/r\overline{W}$) gives an apparently simple but in general rather refractory form (assuming a given set of frequencies of other genes)

$$\varphi(q) = [\overline{W}^{2N}/\sigma^2_{\delta q}][C - 2D\int \overline{W}^{-2N}dq]. \qquad (15)$$

It will be convenient for later reference to cite the less general case $\Delta q = q(1 - q)(s + tq)$, $\sigma^2_q = q(1 - q)/2N$ which allows for any degree of dominance, provided s and t are both small.

$$f(q) = [e^{4Nsq+2Ntq^2}/q(1 - q)][C - 2D\int e^{-(4Nsq+2Ntq^2)}dq]. \qquad (16)$$

Non-stationary States.—The general case, in which the proportion at each value of q is a function of time as well as of q itself, is given by the following, of which equation (1) is the case in which the left-hand member is zero. Time (T) is measured in generations.

$$\frac{\partial\varphi(q, T)}{\partial T} = \frac{1}{2}\frac{\partial^2}{\partial q^2}[\sigma^2_{\delta q}\varphi(q, T)] - \frac{\partial}{\partial q}[\Delta q\varphi(q, T)]. \qquad (17)$$

This can be reduced to an ordinary differential equation in the case in which the distribution has reached stability of form, with all classes (except the terminal ones) falling off at the same rate. Let $K = -\frac{1}{\varphi(q, T)}\frac{\partial\varphi(q, T)}{\partial T}$ be the rate of decay per generation.

$$\frac{1}{2}\frac{d^2}{dq^2}(\sigma^2_{\delta q}\varphi(q)) - \frac{d}{dq}(\Delta q\varphi(q)) + K\varphi(q) = 0. \qquad (18)$$

It may easily be verified that for the case in which fixation is occurring under the uncomplicated effect of inbreeding ($\Delta q = 0$, $K = 1/2N$) the only solution that does not involve negative frequencies is

$$\varphi(q) = 1, \qquad \text{or } f(q, T) = C_0e^{-T/2N}. \qquad (19)$$

In the case of irreversible mutation at an appreciable rate, $\Delta q = v(1 - q)$ the rate of decay is easily shown to be $K = v$. Equation (18) is satisfied by the following value, originally derived by a different method.

$$f(q) = 2vq^{4Nv-1}. \qquad (20)$$

An analogous solution applies to the effect of swamping by immigration from a population in which the gene in question is fixed ($\Delta q = m(1 - q)$, $K = m$)

$$f(q) = 2mq^{4Nm-1}. \qquad (21)$$

Comparison with Results by Other Methods.—The first attempt at determining the distribution of gene frequencies was made by R. A. Fisher[1] who arrived at differential equations for certain special cases, in terms, however, of a different variable than gene frequency, $\vartheta = \cos^{-1}(1 - 2q)$, used in order to make the sampling variance constant. A discrepancy between the rate of decay ($K = 1/4N$), derived by him for the case in which $\Delta q = 0$, and the value, $1/2N$, given by a general method[12, 13] for determining the rate of fixation of genes under any system of mating, led the present author[3, 4] to a different approach. The condition for a stationary state of the intermediate classes except for possible decay at rate K, was represented by the following equation in which q and x are recipient and donor classes, respectively, in the exchanges which occur from one generation to the next.

$$(1 - K)\frac{\varphi(q)}{2N} =$$

$$\frac{(2N)!}{(2Nq)![2N(1-q)]!}\int_0^1(x + \Delta x)^{2Nq}(1 - x - \Delta x)^{2N(1-q)}\varphi(x)dx. \quad (22)$$

It could easily be seen that if $\Delta x = 0$, the equation is satisfied by $\varphi(q) = \varphi(x) = 1$, $K = 1/(2N + 1)$, the latter at least a close approximation to the rate of decay expected in this case. For the simplest stationary state, $K = 0$, $\Delta q \doteq 0$, the expression $\varphi(q) = Aq^{-1} + B(1 - q)^{-1}$ is indicated (cf. 12). Approximate solutions could also readily be obtained for the linear pressures of mutation and migration. Selection presented more difficulty.

On inspection of these results in manuscript, Fisher[2] was able to correct and extend his equations to obtain the following:

I. Case of uniform decay ($\Delta q = 0$)

$$\frac{\partial y}{\partial T} = \frac{1}{4n}\left[\frac{\partial^2 y}{\partial \vartheta^2} + \frac{\partial}{\partial \vartheta}(y \cot \vartheta)\right]. \quad (23)$$

$$y = A_0 e^{-T/2n} \sin \vartheta \qquad \text{(cf. (19))}. \quad (24)$$

II. Stationary state, no selection ($\Delta q \doteq 0$)

$$\frac{dy}{d\vartheta} + y \cot \vartheta = -4nB. \quad (25)$$

$$y = A \operatorname{cosec} \vartheta + 4nB \cot \vartheta \qquad \text{(general, cf. (12))}. \quad (26)$$

$$y = A \operatorname{cosec} \vartheta \qquad \text{(symmetrical case, cf. (14))}. \quad (27)$$

$$y = 4nB(\operatorname{cosec} \vartheta + \cot \vartheta) \qquad \text{(unidirectional mutation, cf. (13))}. \quad (28)$$

III. Stationary state, selection, no dominance, $\Delta q = aq(1-q)$

$$\frac{dy}{d\vartheta} - (2an \sin\vartheta - \cot\vartheta)y = -4anA. \quad (29)$$

$$y = \operatorname{cosec}\vartheta(2A + Be^{-2an\cos\vartheta}) \quad \text{(general, cf. (16), } t = 0). \quad (30)$$

$$y = 4\operatorname{cosec}\vartheta\frac{(1 - e^{-2an(1+\cos\vartheta)})}{(1 - e^{-4an})} \quad \text{(unidirectional mutation).} \quad (31)$$

In cases I and II, the results agreed with those obtained from the integral equation (22), as may be seen by making the substitutions $\cos\vartheta = 1 - 2q$, $yd\vartheta = \varphi(q)dq$, $d\vartheta/dq = 1/\sqrt{q(1-q)}$. In case III it was the author's turn to make a correction in the selection term (published first[3] as e^{2Nsq}), by taking cognizance of a series of small terms erroneously thought to be negligible but which actually doubled the exponent. With this correction, there was agreement.[4]

The most general result[5] obtained for the completely stationary state by solution of (22) took into account all of the factors of change in the form $\Delta q = v(1-q) - uq - m(q - q_t) + q(1-q)(s + tq)$, $\sigma^2_{\delta q} = q(1-q)/2N$.

$$\varphi(q) = Ce^{4Nsq+2Ntq^2}q^{4N(mq_t+v)-1}(1-q)^{4N[m(1-q_t)+u]-1}. \quad (32)$$

This agrees with that obtained by substituting these values of Δq and of $\sigma^2_{\delta q}$ in (6).

The most general result[5,6] obtained by this method for the case of steady flux was for $\Delta q = q(1-q)(s + tq)$.

$$f(q) = [e^{4Nsq+2Ntq^2}/q(1-q)]\,[C - 2Dqe^{-(2Nsq+Ntq^2)}\psi(2Nsq, 2Ntq^2)]. \quad (33)$$

where

$$\psi(a, 0) = 1 + \frac{a^2}{3!} + \frac{a^4}{5!} + \frac{a^6}{7!}\ldots = (e^a - e^{-a})/2a$$

$$\psi(0, b) = 1 + \frac{b}{3!} + \frac{7b^2}{5!} + \frac{27b^3}{7!}\ldots E_m b^m$$

$$E_m = (E_{m-1} + E_{m-2})/2m(m+1).$$

No recurrence formula was recognized for the joint terms, $\psi(2Nsq, 2Ntq^2)$ but the coefficients were calculated[7] up to those pertaining to q^8.

The probability of fixation of a single mutation ($C = 2v$, $D = ve^{2Ns+Nt}/\psi(2Ns, 2Nt)$ for irreversible mutations from class $q = 0$, or $C = 0$, $D = -ve^{-(2Ns+Nt)}/\psi(2Ns, 2Nt)$ for irreversible mutations from class $q = 1$), could be calculated from the ratios of the subterminal classes, (Prob. $= \sqrt{s/2N}$ for a recessive mutation with selective advantage s, Prob. $= 2s$ for a dominant mutation with selective advantage s, or for a semidominant

with selective advantage s in the heterozygote). The last agrees with Fisher's conclusion.[2] Equation (33), with $t = 0$, is indeed equivalent to (31).

Comparison of (33) with (16) shows that if the former is correct, the following must hold:

$$\psi(2Nsq, 2Ntq^2) = (e^{2Nsq+Ntq^2}/q)\int e^{-(4Nsq+2Ntq^2)}dq. \qquad (34)$$

This was tested by expanding the two exponentials in (34), integrating each term of the second one and combining. The coefficients were in all cases identical with those published[7] for $\psi(2Nsq, 2Ntq^2)$.

Equation (22) also gave the solution (20) for the case of uniform decay under an appreciable mutation rate.[4]

The integral equation (22) and the differential equation (18) are clearly equivalent to a close approximation. They are not exact mathematical equivalents, however, as may be seen from the fact that K must be put $1/(2N + 1)$ in (22) if $\Delta x = 0$ to give the solution $\varphi(q) = 1$, while it takes its true value $1/2N$ in (18) to give the same result. In the other cases (except (12)) second order terms have been omitted in the series, obtained as solutions of the integral equation, which do not appear in the solutions of the differential equation. Neither equation, of course, represents the natural conditions exactly since integration is substituted for summation and differentials for minimal steps $(1/2N)$ in gene frequency.

[1] Fisher, R. A., *Proc. Roy. Soc. Edinburgh*, **42**, 321–341 (1922).

[2] Fisher, R. A., *Ibid.*, **50**, 205–220 (1930).

[3] Wright, S., *Amer. Naturalist*, **63**, 556–561 (1929).

[4] Wright, S., *Genetics*, **16**, 97–159 (1931).

[5] Wright, S., these PROCEEDINGS, **23**, 307–320 (1937).

[6] Wright, S., *Ibid.*, **24**, 253–259 (1938).

[7] Wright, S., *Bull. Amer. Math. Soc.*, **48**, 223–246 (1942).

[8] Kolmogorov, A., *C. R. de l'Acad. des Sciences de l' U.R.S.S.*, **3** (7), 129–132 (1935).

[9] Wright, S., these PROCEEDINGS, **24**, 372–377 (1938).

[10] Wahlund, S., *Hereditas*, **11**, 65–106 (1928).

[11] Wright, S., *Genetics*, **24**, 538–552 (1939).

[12] Wright, S., *Ibid.*, **6**, 111–178 (1921).

[13] Wright, S., *Amer. Naturalist*, **61**, 330–338 (1922).

EFFECT OF SMOKING ON TASTE THRESHOLDS FOR PHENYL-THIO-CARBAMIDE (PTC)

By Ada R. Hall and Albert F. Blakeslee

Department of Zoölogy and Physiology, Wellesley College, and Department of Botany, Smith College

Communicated November 2, 1945

In the course of experiments on tasting ability the question arose as to whether a person's threshold was the true one if he had been recently smoking. Salmon and Blakeslee (1935) tested a large group for PTC threshold and recorded the type of smoking for each individual (heavy, moderate, none). These records show that the position of the subject's threshold on the PTC scale is not correlated with his smoking habits. That is, a heavy smoker is just as likely to have a low threshold as a high one and vice versa. The amount that an individual might vary from his resting threshold when he smoked again had not been tested. (Resting threshold is here used as the value obtained after nine or more hours of abstinence from tobacco.)

Such a series of tests was therefore undertaken to show the effect, if any, that tobacco has on the taste apparatus, and to determine how long a time must elapse between smoking and a return to the original threshold.

In a search of the literature for the effect of tobacco on the various senses it was found that the following reactions have been studied: visual acuity, skin pressure sense, eye accommodation, skin reaction to slight electric currents, taste preferences, and mental and physical efficiency. In addition a number of workers have analyzed smoke and the tobacco in its various forms, snuff, pipe cuts, and cigarettes, for the ingredient causing the reactions. Bogen (1936) after careful analyses reports that no single explanation applies to the problem of irritation by tobacco smoke. The innocent bystander receives the sidestream rich in ammonia and other alkaline substances quite irritating to the eye and nasal passages. The smoker who gets the main stream through the cigarette or pipe stem has a more acid material thereby receiving his nicotine (alkaloid) as salts which are less irritating. The smoker does get heat, pyridine, volatile acids, tarry and phenolic constituents, and the aldehydes, furfural and acrolein. These all lead to irritation of the membranes. All workers are fairly well agreed that nicotine in whatever form absorbed is the principal toxic agent wherever the nervous system is involved.

Carefully controlled tests on mental and physical efficiency before and after smoking show a definite correlation between change in efficiency and smoking. Hull (1924) points out the following facts:

1. There is a large and uniform increase in the tremor of the hand lasting an hour and 23 minutes.
2. There is marked and uniform stimulation of the heart which is still present an hour and 40 minutes after smoking.
3. There is a minute increase in the speed of reading reaction time.
4. There is a gain in the rate of complex mental addition lasting an hour and 15 minutes but with no measurable effect on accuracy.
5. There is a high probability of loss in auditory memory and efficiency in rote learning, with recovery in an hour.

In a series of ten mental tests Bush (1914) reports a 10.5% decrease in efficiency after smoking.

Accommodation time, both near-to-far and far-to-near sight, has been studied over a period of years at Wellesley College. Homewood and Howe (1937) have shown that accommodation time is definitely decreased during the first 20 minutes after smoking, but increased from 40 to 60 minutes after. This agrees with Schrumpf-Pierron's (1927) statement that the effect of tobacco on the nervous system is first a stimulation and then a depression. They also found that the occasional smoker was stimulated more than the habitual one and depressed sooner and to a greater extent.

Wenusch and Schöller (1936) working on skin pressure found that sensitivity to both hair pressure and pendulum stroke was changed during smoking. They do not record any stimulating effect of smoking, only a depressant one.

From a survey of the literature on the changes in mental and physical states after smoking, and from a series of electrical tests on the finger before and after smoking, Mendenhall (1930) concludes that tobacco has the same effect as rest. If a person is tired and depressed, smoking will stimulate and bring him back to normal, if overexcited it will quiet him by its depressing action. If he is in a well-rested state it has no marked effect.

Sinnot and Rauth (1937) found the thresholds for sugar and salt high in smokers, but during a period of several days during which six smokers had abstained, their thresholds fell to the level of non-smokers. Laird (1939) tested the effect of the smoking habit on the preference for sweet or tart pineapple juice. He found no difference in the taste preferences of smokers and non-smokers among men, at any age, nor among women up to forty years. Among the women of fifty to sixty-eight years the non-smokers were like the other groups, but the smokers were predominantly in favor of the tart juice rather than the sweet. He did not test each individual before and after smoking, nor does he deal with threshold ability.

In testing the present group of subjects for the effect of smoking on tasting ability each person was his own control. Smokers were tested

only after nine or more hours of abstinence from tobacco. Of the 60 subjects, 32 were from the staff at Cold Spring Harbor and 28 were advanced students and staff at Wellesley College; 32 were habitual smokers and 28 non-smokers; 24 were men and 36 women. Concentrations of PTC were used from 1:5120 M up to 1:312.5 using a factor of two, such that each solution used in a test was twice as concentrated as that previously administered. (M in these numbers represents 1000.) In each test approximately 0.6 cc. was given by means of the straw method.

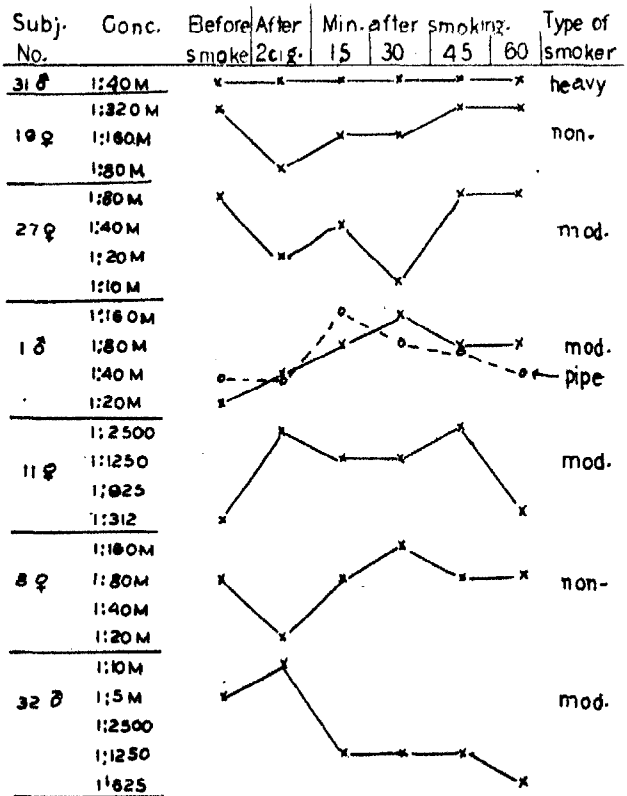


FIG. 1. ABILITY TO TASTE P.T.C. AFTER SMOKING.

Two types of experiments were performed. The procedure for the first series of tests was as follows:

1. The PTC threshold was determined.
2. The subject smoked two cigarettes of a standard brand in 10–15 minutes.
3. The PTC threshold was determined immediately after smoking and

at 15-minute intervals for one to two hours. (In several cases the tests ran still longer.) In the second series 10 individuals of the first series, both smokers and non-smokers, were tested as above except that the smoke was drawn through a dry flask for cooling and into the nose through a sterilized nose-piece keeping the mouth closed so that the smoke did not touch the taste buds.

A summary of the results of the first series may be found in table 1. Figure 1 shows typical PTC curves for each of the behavior groups.

TABLE 1

SUMMARY OF THE REACTIONS TOWARD PHENYL-THIO-CARBAMIDE AFTER SMOKING (SMOKE TAKEN IN THROUGH THE MOUTH)

CHANGE IN TASTING ABILITY	INDIV. IN GROUP	% OF TOTAL TESTED	NO. PTC GRADES		RETURN TO INITIAL LEVEL		SMOKER		SEX	
			INC.	DEC.	NO.	TIME	YES	NO	♂	♀
A. None	4	6.6	...	...	4	No change	2	2	1	3
B. Decreased only	32	53.3	...	1–5	22	30′–60′	15	17	12	21
					10	Over 60′				
C. Increased only	10	16.7	1–4	...	6	15′–60′	8	2	5	5
					4	Over 60′				
D. Decreased then increased	12	20.0	1–2	1–2	7	30′–60′	5	7	5	7
					5	Over 50′				
E. Increased then decreased	2	3.3	1	1–2	2	Over 60′	2	..	1	1
Totals	60	100			35	15′–60′	32	28	24	36
					21	Over 60′				

There were three types of reaction:

1. No change in tasting ability, group A (6.6%).
2. Decrease in tasting ability, groups B and D (73.3%): group B returned to initial level (53.3%); group D returned to initial level and then showed an increase (20%).
3. Increase in testing ability, groups C and E (20%): group C returned to initial level (16.7%); group E returned to initial level and then showed a decrease (3.3%).

These figures check very well with those of Mendenhall (1930) working on sensitivity to electric shock. He found that when tested before and after smoking (750 observations) 72.2% of the cases showed a depression while 28.3% showed a stimulation. In our work for taste 73.3% were depressed, 20% were stimulated and 6.6% showed no change.

Considering the time at which depression and stimulation occur the following summary may be made:

Time at which tasting ability is first lowered (groups B, D):

38 during smoking (10–15 minutes)
4 by 15 minutes after smoking
2 by 30 minutes after smoking

Time at which tasting ability is first increased (groups C, E):

8 during smoking
3 by 15 minutes after smoking
1 by 30 minutes after smoking

Time of greatest increase after an initial decrease (groups B, D):

5 by 15 minutes after smoking
10 by 30 minutes after smoking
16 by 45 minutes after smoking
10 by 60 minutes after smoking
3 not increased again by 60 minutes (one was still low at 105 min.)

Thus it may be seen that the initial depression occurs during smoking or by 15 minutes after. For those who show no initial decrease, stimulation occurs for most individuals during smoking or in 15–30 minutes after. If there is an initial decrease the return to the original threshold or to a period of stimulation is for most subjects at 30–45 minutes after smoking.

The taste apparatus according to Ranson (1939) is composed of taste buds on the tongue from which two sets of special afferent visceral fibers pass to the tractus solitarius via the chorda tympani (of the seventh cranial nerve) and the glosso-pharyngeal nerve. These taste fibers connect with the anterior part of the nucleus of the tractus solitarius. Further connections are rather vague but certainly reflexes to the gustatory center and the motor nuclei for mastication and swallowing exist. A cerebral center is not definitely fixed but a spot near the anterior end of the temporal lobe is thought to have such function. There are at least three synapses then between the surface of the tongue and consciousness.

According to the work done on other sensory-motor arcs the expected curve for taste should show an initial level, a more sensitive period, and then a depression period. But we have here an added factor to complicate results, namely, the exposure of the sensory endings, the taste buds, to the various drugs resulting from the combustion. It may be that the depressant effect on the taste buds antedates or coincides with the stimulative effect on the nerve cells and thus we get several forms of curves for the first half hour depending on the comparative strength of the two factors. Some individuals keep to the same level at first with later depression, others have a sharp early depression followed by stimulation, while others are stimulated immediately and to a greater degree than the depression

of the taste buds. We felt that it might be enlightening so to conduct the smoking period that the taste buds would not be touched by the smoke. Of the 10 individuals so tested 3 showed no change in tasting ability for 30 minutes to an hour followed then by a depression. Six had an initial stimulation followed at 30 to 45 minutes by a depression. These persons had all shown initial depression when smoking by mouth. One subject (of group C) who showed only stimulation when smoking by mouth had the same type of reaction (stimulation only) when smoking by "nose." Figure 2 shows the mouth and "nose" curves for two of these subjects.

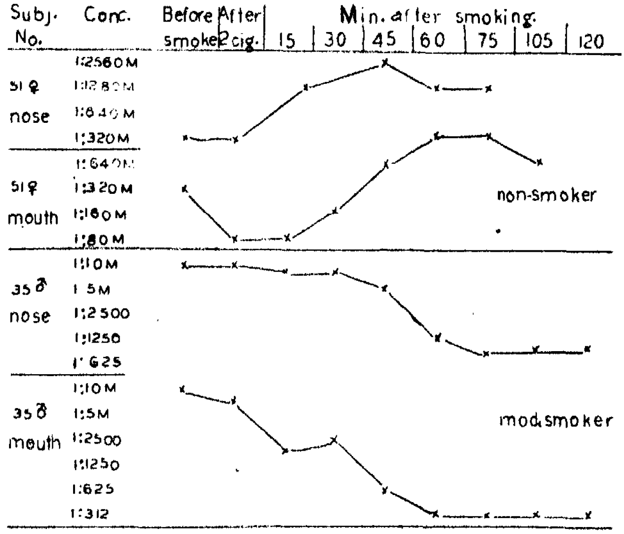


FIG. 2. EFFECT OF SMOKING BY NOSE.

It would seem then that the sharp depression noted for the larger number of subjects in series 1 may be due to the direct action of some product of combustion on the taste buds rather than to the effect of nicotine on the nerve cells.

We may therefore make the following conclusions:

1. There is a definite effect on the individual's ability to taste PTC after smoking, 73.3% of the subjects requiring stronger solutions in order to taste after smoking and 20% of them tasting weaker solutions.

2. The time which must elapse after smoking before the individual is a proper subject for tasting experiments varies with the individual. In the present series of tests only 58% had returned to the resting threshold within an hour, while some took several hours for recovery.

3. The initial effect of smoking in the larger number of cases is a direct dulling of the taste buds by some product of the combustion.

4. The true effect of nicotine on the nerve apparatus for taste appears to be the same as for other nerves tested—an initial stimulation with later depression.

Bogen, E., "Irritant Factors in Tobacco Smoke," *Cal. and West. Med.*, **45**, 342–346 (1936).

Bush, A. D., "Tobacco Smoking and Mental Efficiency," *N. Y. Med. Jour.*, **99**, 519–527 (1914).

Homewood, Jean, and Howe, E. C., "Delayed Effects of Smoking on Visual Accommodation," unpublished thesis of Dept. of Hygiene and Physical Education, Wellesley College, 1937.

Hull, C. L., "The Influence of Tobacco Smoking on Mental and Motor Efficiency," *Psychol. Monog.*, **33**, no. 3, 1–159 (1924).

Laird, D. A., "Effect of Smoking on Taste Preferences," *Med. Rec.*, **149**, 404 (1939).

Mendenhall, W. L., *Tobacco*, Harvard University Press, 1930.

Ranson, W. S., *The Anatomy of the Nervous System*, W. B. Saunders Co., 1939.

Salmon, T. N., and Blakeslee, A. F., "Genetics of Sensory Thresholds: Variations within Single Individuals in Taste Sensitivity for PTC," these PROCEEDINGS, **21**, 78–83 (1935).

Sinnot, J. J., and Rauth, J. E., "Effect of Smoking on Taste Thresholds," *Jour. Gen. Psychol.*, **17**, 151–153 (1937).

Schrumpf-Pierron, Pierre, *Tobacco and Physical Efficiency*, Paul Hoeber Inc., New York, N. Y., 1927.

Wenusch, A., and Schöller, R., "Uber den Einfluss des Rauchens auf die Reizschwelle des Drucksinnes," *Med. Klin.*, **32**, 356–358 (1936).

RESPIRATORY ENZYMES IN PARAMECIUM: 1. CYTOCHROME OXIDASE[*]

BY EDGAR J. BOELL

OSBORN ZOÖLOGICAL LABORATORY, YALE UNIVERSITY

Communicated October 15, 1945

It has been generally stated, as the result of the work of a number of investigators,[1, 2, 3] that cyanide is without effect on the respiratory activity of Paramecium. The report of Kalmus[4] that respiration of Paramecium is depressed by cyanide has been largely disregarded because of serious defects in his experimental technique (cf. Howland and Bernstein[5]). Cyanide was likewise shown to be without effect on the respiration of a number of other ciliates,[6, 7, 8] and the conclusion was reached, as summarized by Lwoff,[8] that "Cette insensibilité à HCN et à CO n'est pas générale chez les Protozoaires, mais est jusqu'ici particulière aux Infusoires." The repeated failure to obtain depression of respiration in ciliates with cyanide and similar inhibitors led naturally to the belief that oxidations in these forms were mediated by a mechanism which was different from that of most aerobic cells.[9]

More recently, a reinvestigation of the effect of cyanide on the respiration of certain ciliates has been made. In these studies a number of technical improvements, not alone in the use of cyanide but also in rearing the animals and preparing them for respiratory measurements, were employed. Thus Baker and Baumberger[10] and Hall[11] showed that respiration in *Tetrahymena geleii* and *Colpidium campylum* could be inhibited by cyanide. Kitching,[12] in a study of contractile vacuole activity, obtained indirect evidence for cyanide susceptibility in the peritrich, Zoothamnion. Finally, the respiration of *Paramecium calkinsi* was shown to be inhibited to approximately 50 per cent of the normal by cyanide (Boell[13] and recently confirmed by Pace[14]) and by azide of the same concentrations as those usually employed for respiratory studies.

Since these inhibitors have been widely used to indicate the operation of the so-called cytochrome-cytochrome oxidase system in normal respiration, the conclusion was drawn[13] that in so far as susceptibility to cyanide and azide can be used as tests, the results suggest that the respiratory mechanism of Paramecium resembles that of most aerobic plant and animal cells. The purpose of this paper is to provide further support for the above conclusion by presenting direct evidence for the occurrence in Paramecium of an oxidative enzyme which presumably is cytochrome oxidase. Apparently this represents the first time that an oxidative enzyme has been obtained from a ciliate and studied by quantitative techniques.

Technique.—Respiratory measurements and cytochrome oxidase determinations were carried out in the Cartesian diver apparatus on known numbers of *Paramecium calkinsi* which had been reared and prepared for study by methods previously described by Boell and Woodruff.[15] In the cytochrome oxidase determinations, highly concentrated suspensions of well-washed Paramecium were used. Duplicate 10-cu. mm. samples were withdrawn in order to obtain a count of the number of animals present in a given volume of medium, and the remainder of the suspension (approximately 0.1 cc.) was ground into a cell-free brei by means of a motor-driven homogenizer. Appropriate amounts of the enzyme preparation were then added to measured quantities of cytochrome-c, ascorbic acid (Merck's Cebione) which had been neutralized with NaOH, and phosphate buffer at pH 7.2. A series of divers was then prepared, each containing a 1.5-cu. mm. sample of the mixture or the same volume of a preparation containing enzyme which had been heated in a boiling water bath for 5 to 10 minutes prior to mixing with cytochrome and ascorbic acid. In some cases divers were prepared so that comparisons of enzyme activity in the presence and absence of added cytochrome could be made. All experiments were conducted at 25°C.

Results.—*1. Inhibition of Respiration by Cyanide and Azide.*—The

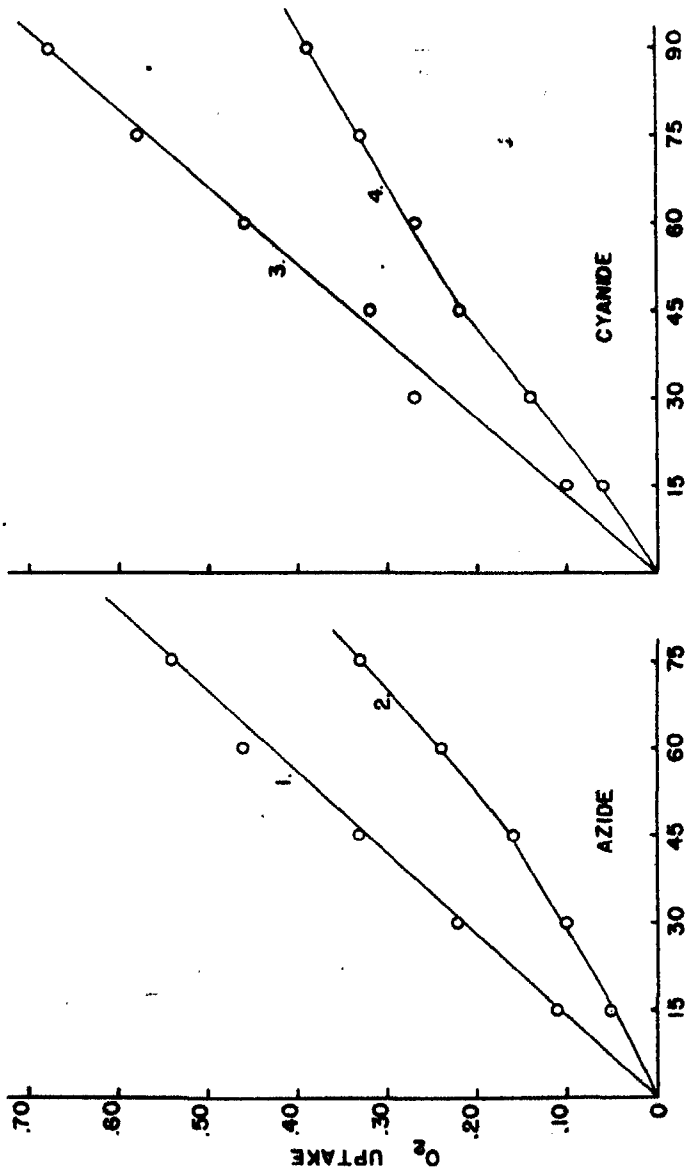


FIGURE 1

The effect of 10^{-2} molar sodium azide (Curve 2) and 10^{-3} molar KCN (Curve 4) on the respiration of *Paramecium calkinsi*. Curves 1 and 3 denote respiratory rates of untreated animals. The pH of the medium with azide was 6.2; that with cyanide, 7.0. Cyanide had been neutralized with HCl before use. An appropriate KCN-KOH mixture[23] was used to absorb carbon dioxide. The ordinate denotes oxygen uptake in mμl. per Paramecium. (One mμl. = 10^{-3} cu. mm.)

inhibition of respiration of intact cells by cyanide, azide, carbon monoxide, etc., has generally been interpreted as due to interference with the functioning of cytochrome and cytochrome oxidase. Thus, susceptibility to cyanide has been used as an indicator of the operation *in vivo* of this system of respiratory catalysts. The results of a typical experiment with these

inhibitors are shown graphically in figure 1. Curves 2 and 4 of this figure show, respectively, the effect of sodium azide and potassium cyanide, each in a final concentration of 10^{-2} molar, on the oxygen consumption of Paramecium. With both these substances respiration is depressed to approximately 50 per cent of the normal, and data are available which show that this inhibition is completely reversible. At acid pH values, however, Paramecia are irreversibly affected by strong concentrations of azide. Such animals apparently lose their ability to maintain water balance, for the contractile vacuoles stop beating and become enormously swollen.

In some experiments in which azide was used, there was a period of accelerated oxygen consumption following the inhibition of respiration. In Curve 2, representing only the first 75 minutes of a longer experiment, such a rise in respiratory rate can be seen. The rate of respiration stimulated by azide may, under certain experimental conditions, considerably exceed the normal, and it is of interest to note that the acceleration of oxygen consumption produced by azide is susceptible to cyanide.

The control respiratory rates, as shown by Curves 1 and 2 in figure 1, although identical, were obtained from two different samples of Paramecium and at different pH values, and, incidentally, it may be mentioned that the average respiratory rate for these animals is essentially identical with that reported for *P. calkinsi* of the same race and mating type by Boell and Woodruff.[15]

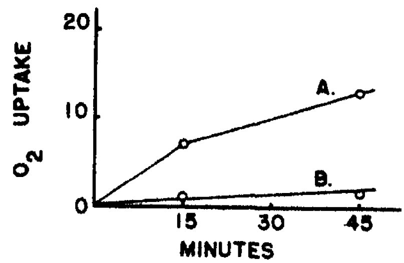


FIGURE 2

Oxidation of *p*-phenylenediamine by oxidase from *Paramecium calkinsi*. Experimental conditions: 25°C.; pH 7.2; final concentrations of cytochrome-c = 1×10^{-4} molar, of *p*-phenylenediamine = 0.033 molar. Oxidase preparation equivalent to approximately 75 Paramecia.

Curve *A* = oxidase + cytochrome + *p*-phenylenediamine; Curve *B* = same but with boiled oxidase. Ordinate denotes oxygen uptake in mμl.

2. *Cytochrome Oxidase Activity.*—Preliminary tests of the cytochrome oxidase activity of Paramecium were performed with 0.033 molar *p*-phenylene diamine as substrate (Fig. 2). However, this compound, although it was readily oxidized, appeared to be toxic to the oxidase as evidenced by a rapid decline in the rate of O_2 uptake (cf. Schneider and Potter[16]). Its use was therefore abandoned. Ascorbic acid in a final concentration of 0.01 or 0.02 molar was used instead, since a number of workers had shown that it can reduce cytochrome-c and could, accordingly, be used as a substrate in tests for cytochrome oxidase.[16–19]

The results of a typical experiment are presented graphically in figure 3 and show unmistakably the presence of an enzyme in Paramecium capable

of oxidizing ascorbic acid. Moreover, since the catalytic oxidation of ascorbic acid is much greater in the presence of cytochrome than in its absence, it seems safe to conclude that the enzyme in this case is cytochrome oxidase. In this connection it may be recalled that Sato and Tamiya[20] reported finding the absorption bands of reduced cytochrome in Paramecium.

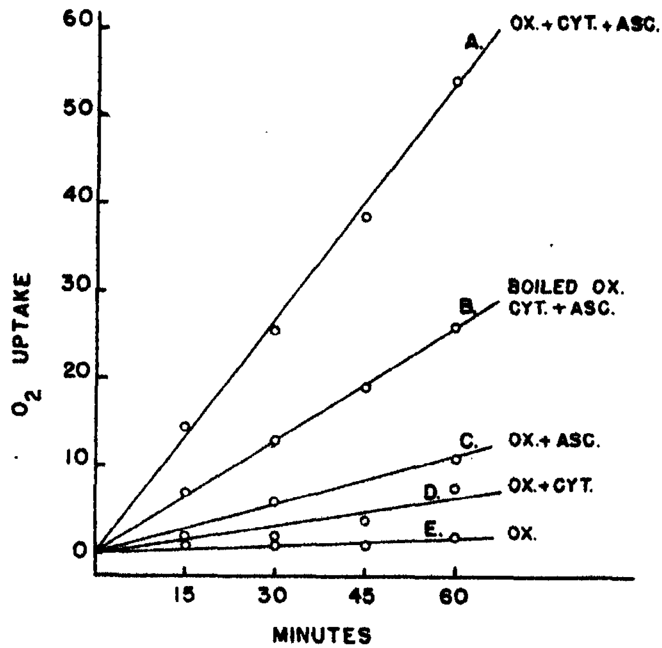


FIGURE 3

Oxidation of ascorbic acid by oxidase from *Paramecium calkinsi*. Experimental conditions: the same as stated for figure 2 except that ascorbic acid in a final concentration of 0.01 molar was used as substrate in place of *p*-phenylenediamine.

Curve *A* = oxidase + cytochrome + ascorbic acid; Curve *B* = same but with boiled oxidase; Curve *C* = oxidase + ascorbic acid; Curve *D* = oxidase + cytochrome; Curve *E* = oxidase. Ordinate denotes oxygen uptake in mμl.

Ascorbic acid in solution shows a fairly rapid oxygen uptake even in the absence of enzyme which has been shown to be due to the presence in the reaction mixture of metallic copper impurities.[19, 20] It is necessary, therefore, to run a blank containing boiled enzyme, cytochrome and substrate in order to determine the non-enzymic oxidation of ascorbic acid. However, since Stotz, *et al.*,[19] have shown that the catalytic effect of metallic copper is eliminated by the presence in the reaction mixture of undenatured

protein, as well as other organic compounds,[21] e.g., glycine,[22] it seems fairly certain that this method introduces an over-correction. Consequently the cytochrome oxidase values obtained are minimal.

Schneider and Potter[16] have described a method of correction which involves making determinations of the oxidase activity at three different concentrations and then extrapolating these to zero in order to determine the oxygen uptake due to non-enzymic catalysis. Such a procedure has been used in the preparation of table 1 which shows, in addition, a direct relationship between enzyme activity and enzyme concentration.

TABLE 1

NUMBER OF ANIMALS IN 1.5 CU. MM. OF REACTION MIXTURE	O_2 UPTAKE mμl/HOUR	CORRECTED CYTOCHROME OXIDASE ACTIVITY	RATIO: ACTIVITY/NUMBER
0	35*	0	..
102	57	22	21
204	76	41	20
306	96	61	20

* Derived by extrapolating the O_2 uptake at other values to zero enzyme concentration.[16]

Summary.—By a micromanometric method, the respiration of *Paramecium calkinsi* has been found to be sensitive to azide and cyanide. In addition, an enzyme capable of oxidizing ascorbic acid and *p*-phenylenediamine in the presence of cytochrome-c and which may, therefore, be designated as cytochrome oxidase has been demonstrated.

[1] Lund, E. J., *Amer. Jour. Physiol.*, **47**, 167–177 (1918).
[2] Shoup, C. A., and Boykin, J. T., *Jour. Gen. Physiol.*, **15**, 107–118 (1931).
[3] Gerard, R. W., and Hyman, L. H., *Amer. Jour. Physiol.*, **97**, 524–525 (1931).
[4] Kalmus, H., *Zs. f. vergl. Physiol.*, **7**, 314–322 (1928).
[5] Howland, R. B., and Bernstein, A., *Jour. Gen. Physiol.*, **14**, 339–348 (1931).
[6] Peters, R. A., *Jour. Physiol.*, **68**, ii–iii (1929).
[7] Pitts, R. F., *Proc. Soc. Exp. Biol. and Med.*, **29**, 542–544 (1932).
[8] Lwoff, M., *C. R. Soc. Biol.*, **115**, 237–240 (1934).
[9] Jahn, T. L., in *Protozoa in Biological Research*, Columbia University Press, New York, 1941.
[10] Baker, E. G. S., and Baumberger, J. P., *Jour. Cell. and Comp. Physiol.*, **17**, 285–303 (1941).
[11] Hall, R. H., *Physiol. Zool.*, **14**, 193–208 (1941).
[12] Kitching, J. A., *Jour. Exp. Biol.*, **15**, 143–151 (1938).
[13] Boell, E. J., *Anat. Rec.*, **84**, 493–494 (1942).
[14] Pace, D. M., *Biol. Bull.*, **89**, 76–83 (1945).
[15] Boell, E. J., and Woodruff, L. L., *Jour. Exp. Zool.*, **87**, 385–402 (1941).
[16] Schneider, W. C., and Potter, V. R., *Jour. Biol. Chem.*, **149**, 217–227 (1943).
[17] Szent-Györgyi, A., *Biochem. Jour.*, **22**, 1387–1409 (1928).
[18] Keilin, D., and Hartree, E. F., *Proc. Roy. Soc. of London*, (B) **125**, 171–186 (1938).
[19] Stotz, E., Harrer, C. J., Schultze, M. O., and King, C. G., *Jour. Biol. Chem.*, **122**, 407–418 (1938).

[10] Sato, T., and Tamiya, H., *Cytologia:* Fujii Jubilee Vol., 1133–1138 (1937).

[11] Barron, E. S. G., DeMeio, R. H., and Klemperer, F., *Jour. Biol. Chem.*, **112**, 625–640 (1935).

[12] Stotz, E., *Ibid.*, **133**, c (abstract) (1940).

[13] Krebs, H., *Biochem. Jour.*, **29**, 1620–1644 (1935).

*OFFSPRING FROM UNBORN MOTHERS**

BY W. L. RUSSELL AHD PATRICIA M. DOUGLASS

ROSCOE B. JACKSON MEMORIAL LABORATORY, BAR HARBOR, MAINE

Communicated October 29, 1945.

The preliminary experiments reported here show that mouse ovaries transplanted from embryos to adult females can develop to maturity and produce offspring.

Eight such operations were performed, three of which have proved successful. The donor ovaries were obtained from embryos aged $13^1/_2$ days (in 3 of the operations) and $14^1/_2$ days (5 operations) post coitus, matings being timed by the vaginal plug method. They were transplanted to the ovarian capsules of young adult animals whose ovaries had been removed immediately before. A male was placed with each host a few days after the operation. Other transplantation experiments have shown that ovarian regeneration frequently occurs in host animals.[1] In order to distinguish offspring of the transplanted ovaries from offspring of regenerated ovaries a mating plan was used which has been described by Russell and Hurst.[1] This plan uses an inbred strain, stock 129, which carries both chinchilla, c^{ch}, and albino, c^a, genes. The embryo donors were obtained from matings of stock 129 albinos and were, therefore, of the genotype c^ac^a. The hosts were hybrids, although for the purpose of this experiment, transplantations could have been made within the 129 stock. The hybrids were the offspring of matings of 129 $c^{ch}c^{ch}$ × C57 black strain and, therefore, had the genetic constitution Cc^{ch}. These Cc^{ch} hosts with c^ac^a ovary implants were mated with 129 $c^{ch}c^{ch}$ males. Offspring of regenerated host ovaries would, therefore, be either Cc^{ch} or $c^{ch}c^{ch}$, while offspring from the transplanted ovaries would be $c^{ch}c^a$.

The animals used have not yet reached the end of their breeding period, but the three successful operations have so far yielded, respectively, 3, 4 and 10 offspring from the transplanted ovaries. In these operations the donor ovaries were obtained from embryos aged $14^1/_2$ days pc and were transplanted to hosts 40 days old. Regeneration of host ovaries occurred to some extent in all three cases as well as in four of the unsuccessful operations. The remaining unsuccessful operation produced no offspring.

The demonstration that offspring can be obtained from transplanted embryonic ovaries, that is, from unborn mothers, introduces a method that should prove of value in several different fields of experimental zoölogy. Some of the possible uses are discussed below.

1. Several dominant gene mutations apparently produce, in the homozygous mutant type, a lethal condition causing death of the animal while still within the uterus of the mother. The evidence rests on the proportion and nature of the abnormal embryos produced by matings of the heterozygotes. Transplantation of ovaries of the lethal type should, first of all, prove whether the ovary can survive in a normal host and, second, if the ovary does function, it would make possible a direct breeding test of the genetics of the lethal type itself. One experiment of this kind is being made by the authors.

2. In the study of chromosome translocations, experiments similar to the above could be undertaken to determine the chromosome make-up of embryos that die presumably from chromosome unbalance.

3. Ovarian transplantation could be used to identify the genetic type of embryos that have to be killed for study. An example will perhaps make this clearer. Mice homozygous for the gene *W* die shortly after birth from severe anemia and can, therefore, be obtained only from matings of the heterozygotes. Grüneberg[2] points out that since the anemia of *WW* embryos is already well marked on the 16th day it must have started earlier. Its actual onset has not been determined because there was no way of identifying the homozygous segregants before that age.

4. Many transplantation experiments that were possible with older ovaries could employ embryonic ovaries to obtain additional information. For example, in studies of differences in maternal environment the foster mother's influence could be brought into play at an earlier stage in the development of the ova.

5. The relation of hormones and other factors to the development and functioning of the ovary could be investigated by transplanting embryonic ovaries to hosts of various ages or to hosts subjected to various experimental treatments.

6. It is possible that transplantation would prove successful with still earlier embryonic stages. In that case it might be a useful technique for investigations on the origin of the germ cells.

7. There is no reason to suppose that the method could not be repeated for any number of successive generations, thus leading to an indefinite number of unborn direct female ancestors. This possibility should be of interest to investigators in several fields of research. In studies of the mammary tumor agent, for example, mice could be obtained whose female ancestors, for any number of generations, had not been nursed.

Summary.—Offspring were obtained from transplanted embryonic

mouse ovaries. This introduces a method that could be used for the genetic identification of some lethals and certain types of embryos that have to be killed for examination. It should prove of value in studies on differences in maternal environments, in experiments on the relation of hormonal and other factors to the development and functioning of the ovary and possibly for investigations on the origin of the germ cells. An extension of the method over successive generations, to produce animals descended from any number of unborn direct female ancestors, would provide still further possibilities for research.

* Aided by a grant from the Rockefeller Foundation and by grants to the Roscoe B. Jackson Memorial Laboratory from the National Cancer Institute, the Jane Coffin Childs Memorial Fund for Medical Research, the Anna Fuller Fund and the International Cancer Research Foundation.

[1] Russell, W. L., and Hurst, J. G., these PROCEEDINGS, 31, 267–273 (1945).

[2] Grüneberg, H., *The Genetics of the Mouse*, Cambridge Univ. Press, 1943, 412 pp.

---

*INDUCTION OF CONJUGATION IN PARAMECIUM BURSARIA AMONG ANIMALS OF ONE MATING TYPE BY FLUID FROM ANOTHER MATING TYPE**

BY TZE-TUAN CHEN

DEPARTMENT OF ZOÖLOGY, UNIVERSITY OF CALIFORNIA, LOS ANGELES

Communicated October 25, 1945

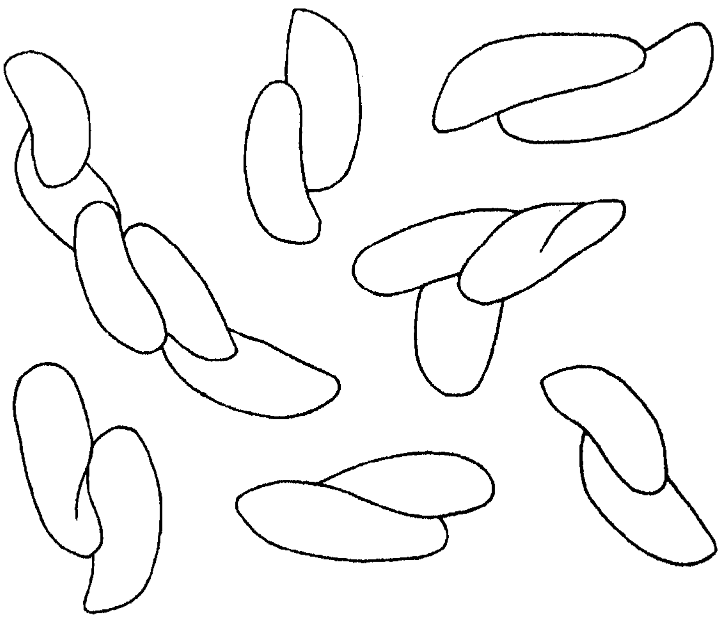
In these experiments, it was found that animal-free fluid from a Russian clone (Ru22) induces clotting and conjugation among animals of another mating type, even though the latter belongs to a different variety.[1]

Most of the present work was done on the effects of fluid from cultures of a Russian clone (Ru22) on animals of mating types belonging to four other varieties. This Russian clone (together with a number of other Russian clones) was sent to the present writer from Moscow by Professor G. F. Gause. Jennings and Opitz[2] found that this clone did not mate with any of the four varieties known at that time and they were of the opinion that this clone possibly belongs to a fifth variety. In the present paper it is called the fifth variety. (Apparently only one mating type of this variety has been found.)

The animals of this Russian clone (as well as those of other clones referred to in this paper) were cultured in essentially the manner described by Jennings.[3]

In testing the effect of fluid from the Russian clone on animals of another mating type, the fluid is first taken out of the culture (Ru22) with a micro-

pipette; it is then placed in depression slides and carefully examined under a dissecting microscope. All the animals, if any are accidentally included in the fluid, are removed. Then a number of animals of the clone to be tested, together with a very small quantity of fluid from this clone, are added to the Ru22 fluid. The depression slides are then left in a moist chamber and examined from time to time. As a control, fluid from the clone to be tested is first taken out of the culture, placed in a depression slide, and then animals from the same culture (or another culture of the same clone) are added.



1a

FIGURE 1a

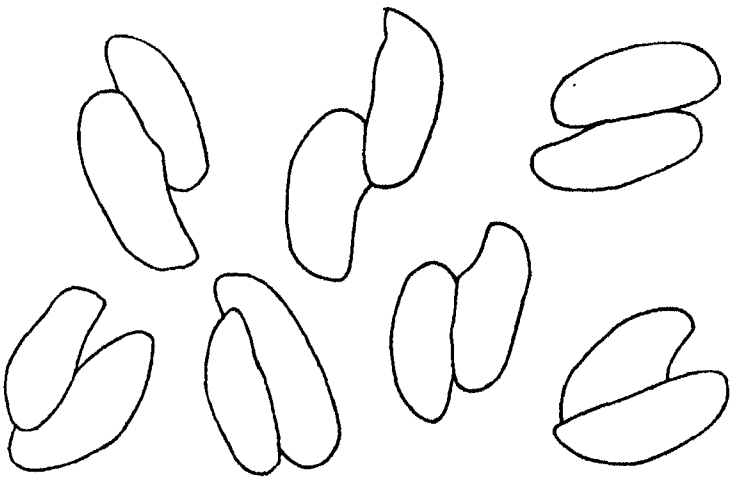
Clotting and pair-formation among animals of an English clone (En1) induced by fluid from a Russian clone (Ru22).

The fluid from cultures of the Russian clone (Ru22) was tested on animals of a number of mating types belonging to four other varieties (II, III, IV, VI) and especially on animals of an English clone (En1) belonging to the sixth variety.

Effects of the Ru22 Fluid on Animals of the English Clone (En1).—This English clone (En1) was collected by Professor E. G. Pringsheim in Cambridge, England, and was sent by him to the present writer, who subsequently made numerous tests on it. The results so far obtained show that it will not mate with animals of any of the known varieties but will con-

jugate readily with animals of a clone collected from Prague, Czechoslovakia (also sent by Professor Pringsheim). It seems clear that this English clone together with the Czechoslovakian clone constitute a new variety, which is designated as Variety VI.

Clotting: The animal-free fluid from Ru22 cultures apparently renders the surface of the animals of the English clone sticky so that the animals adhere to one another, forming clots and pairs (Fig. 1*a*). No such clotting or pair-formation was ever observed in the controls. The clotting or group-formation of animals of the English clone (one mating type)[4] induced by the fluid of the Russian clone (another mating type) differs somewhat from the clotting which occurs when animals of diverse mating types are mixed.



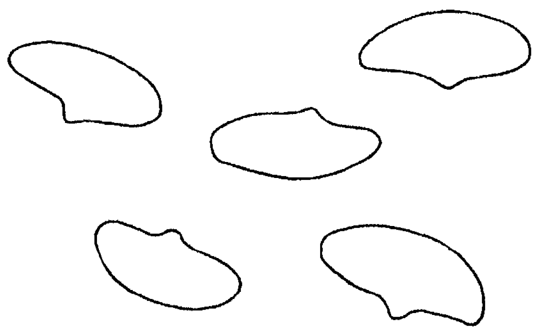
Ib

FIGURE 1*b*

Pairs formed among animals of an English clone (En1) induced by the fluid from a Russian clone (Ru22).

The difference is twofold: (1) The clotting among animals of the same mating type occurs much later than that which usually occurs when animals of diverse mating types are mixed. In the former case, agglutination occurs 4–6 (usually more) hours after the animals of one mating type are introduced into the fluid of another mating type; in the latter case, agglutination occurs immediately or almost immediately. (2) The clots formed by animals of the same mating type are much smaller than those formed as a result of mixing animals of diverse mating types. In induced group-formation, each clot usually consists of 2–7 individuals. Clots formed as a result of mixing animals of diverse mating types are usually larger; each of the larger clots often consists of many individuals.

Conjugation: Pair-formation among animals of the same mating type[5] differs somewhat from pair-formation when those of diverse mating types are mixed. In the first place, in induced conjugation, the pairs are formed much later (4–6 or more hours after the animals are introduced into the fluid of another mating type). When animals of diverse mating types are mixed, usually many pairs are formed in one or two hours. Secondly, in induced conjugation many of the pairs are atypical. Although some of the pairs are hardly distinguishable from those formed as a result of mixing animals of diverse mating types, other pairs are atypical in that the two conjugants do not have the relative positions typical of normal conjugating pairs (Fig. 1*b*). Thirdly, in many pairs, the conjugants are not held as firmly together as those formed in ordinary conjugation. As a result, in a number of cases the conjugants become separated when fixed. This is true even after the animals have been conjugating for more than 20 hours.



1c

FIGURE 1*c*

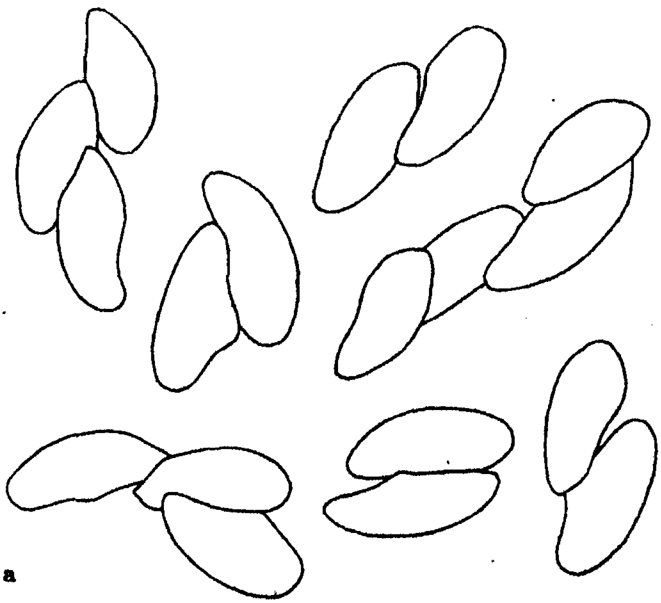
Single animals of an English clone (En1) undergoing perhaps autogamy or endomixis or some other type of nuclear processes induced by the fluid from a Russian clone (Ru22).

Fixed and stained preparations of induced pairs show that nuclear changes occur in the conjugants and anlagen are formed in the ex-conjugants. The nuclear changes in induced conjugation are being studied and will be reported on separately.

Induction of Nuclear Changes in Single Animals: The Ru22 fluid induces not only clotting and conjugation, as already described, but also nuclear changes in some individuals not associated with any others. These single animals develop a paroral cone (Fig. 1*c*). This structure has been reported in *Paramecium aurelia* undergoing autogamy (endomixis).[6] Preliminary cytological studies show that nuclear changes take place in these single animals. It remains to be determined whether they undergo autog-

amy, or endomixis, or some other type of nuclear processes. These single animals may be found together with the conjugating pairs; at other times, however, they may be found alone.

In addition to the induction of clotting and conjugation among animals of the same mating type and of nuclear changes in some single animals, there are other effects of the Ru22 fluid on the members of the English clone. These additional effects are recognizable even before clotting and conjugation occur. The animals become (1) sluggish in movement, (2) darker in color, and (3) distorted in form to a greater or less extent. No such changes were found among the animals in the controls.



2a

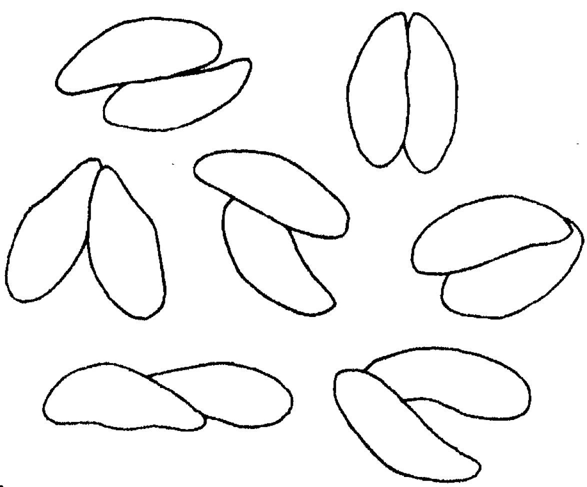
FIGURE 2a

Clotting and pair-formation among animals of an American clone, MM17 (Variety II), induced by fluid from a Russian clone (Ru22).

It should be noted that only a percentage (usually small) of the animals are visibly affected by the Ru22 fluid, and hence only a small number of them undergo clotting and conjugation; while a few solitary animals undergo nuclear changes. It is difficult to understand why only a percentage of animals are obviously affected. It seems probable that there are differences in physiological condition even among the animals of the same culture. It thus seems that those individuals which are affected differ from others added to the fluid at the same time.

Effect of Washing Animals.—In order to determine whether or not induction of clotting and pair-formation among members of the English clone En1 by the Ru22 fluid might be due to some impurities in the cultures of the Russian clone, on November 23, 1944, a number of the animals of this Russian clone were carefully washed and then recultured. Some weeks later the fluid was tested again on the animals belonging to the English clone En1. Clotting and pair-formation were again observed, as described above.

Effects of the Ru22 Fluid on Animals Belonging to Some Additional Varieties.—The fluid of the Russian clone Ru22 is apparently effective in inducing conjugation throughout the species since it induces conjugation in all the four varieties so far tested. The one other variety (I) is being tested.



2b

FIGURE 2b

Pairs formed among animals of an American clone, MM17 (Variety II), induced by the fluid from a Russian clone (Ru22).

In addition to inducing conjugation among the animals of the English clone (En1) belonging to the sixth variety, this Ru22 fluid also induces clotting and conjugation in each of the following clones: MM17 (mating type F) and Gr13 (type K) of the second variety, clones Gr1 (type N) and Pi3 (type O) of the third variety, and clone Ru3 (type R) of the fourth variety.[7] Clotting and pair-formation in these varieties are essentially the same as those which occur among animals of the English clone as described above (cf. Figs. 1*a*, 1*b* and 2*a*, 2*b*). No clotting or conjugation was ever observed in the controls. Some other clones of these three varieties were

also tested but as yet no clotting or conjugation has been observed. These negative results might be due to differential effects of the Ru22 fluid on different clones of the same variety.

Discussion.—The phenomena in *P. bursaria* described in this paper are possibly related to those of "sex stuffs" found in certain algae (Geitler,[8] Moewus[9]) and Protozoa (Kimball[10]). In *Tetraspora lubrica* Geitler found that centrifugates of culture fluid of one sex cause atypical group-formation among cells of the other sex. Copulation, however, does not follow this type of group-formation. In *Chlamydomonas eugametos* Moewus discovered the same effect of filtrates or centrifugates of culture fluid of one sex on individuals of the other sex. In *Euplotes patella* Kimball found that conjugation can be induced between animals of the same mating type by placing them in fluid in which another mating type has been living.

Another possibility is that the phenomena in *P. bursaria* described in this paper are similar to those of the "killers" as reported by Sonneborn[11] for *P. aurelia* in view of the following facts: (1) The Ru22 fluid makes the animals of the English clone En1 sluggish, darker and distorted in shape and these phenomena are similar to those in lethal interactions between diverse stocks in *P. aurelia*. (2) In *P. aurelia* Sonneborn (unpublished) found that one of the "killers," among other effects, induced pairing as described in this paper for *P. bursaria*.

* Aided by grants from the Committee for Research in Problems of Sex of the National Research Council and from the University of California.

[1] The fluid of two other Russian clones (Ru21, Ru30) appears to have the same effects as those of Ru22 fluid. These three clones belong to the same mating type and are characterized by a large body and an unusually large micronucleus. These clones are vigorous; flourishing cultures can be easily obtained.

[2] Jennings, H. S., and Opitz, P., *Genetics*, **29**, 576–583 (1944).

[3] Jennings, H. S., *Ibid.*, **24**, 202–233 (1939).

[4] Of course, it is also possible, though not probable, that the Ru22 fluid causes a change in mating type in some of the animals of the English clone, and hence in this case it is merely a conjugation between animals of diverse mating types.

[5] In addition to the conjugating pairs, some "threes" were also observed. (A group of three animals in conjugation is to be called a "three.") In each "three," nuclear changes occur in all three conjugants and anlagen are formed in the ex-conjugants.

[6] Diller, W. F., *Jour. Morph.*, **59**, 11–67 (1936); Caldwell, L., *Jour. Exper. Zool.*, **66**, 371–407 (1933).

[7] As yet only a few clones of these three varieties have been tested. It is likely that induced conjugation will be found in more clones when a greater number of clones are tested.

[8] Geitler, L., *Biol. Zentralbl.*, **51**, 173–187 (1931).

[9] Moewus, F., *Arch. Protistenk.*, **80**, 469–526 (1933).

[10] Kimball, R. F., *Genetics*, **27**, 269–285 (1942).

[11] Sonneborn, T. M., *Amer. Nat.*, **73**, 390–413 (1939).

MARKOFF CHAINS WITH REVERSE TRANSITIONS

By I. OPATOWSKI

THE UNIVERSITY OF CHICAGO

Communicated October 15, 1945

Consider a system which may take $n + 1$ states: 0, 1, ..., n. Let the probability of transitions $(i - 1 \rightarrow i)$ and $(i + 1 \rightarrow i)$ during any time dt be, respectively, $k_i dt + o(dt)$ and $g_i dt + o(dt)$. Let these be the only transitions possible during dt and let the system be in the state 0 at $t = 0$. If $Y_i(t)$ is the probability that the system be in the state i at the time t, then $Y_0(0) = 1$, $Y_i(0) = 0$ for $i \geqslant 1$ and[1]

$$dY_i/dt = k_i Y_{i-1} - k_{i+1} Y_i + g_i Y_{i+1} - g_{i-1} Y_i, \tag{1}$$

where by definition $Y_{n+1} = Y_j = g_j = 0$ if $j < 0$. In certain biophysical applications[2] of (1) the only function of interest is $Y_n(t)$ and $k_i = k$ or $= 0$, $g_i = g$ or $= 0$, where k, g are constants > 0. The purpose of this paper is to derive, under these assumptions, an explicit expression of $Y_n(t)$. The method used is that of Laplace transformation which has been introduced into this field by H. Bateman.[3] This transformation seems to be essential here because it shows an intimate connection of our problem with *Tchebychef polynomials.*

Put $y_i(s) = \int_0^\infty e^{-st} Y_i(t)dt$ and consider first the following three processes.

First Process.—Let $k_i = k(1 \leq i \leq n)$, $g_i = g(0 \leq i \leq n - 1)$, $k_{n+1} = 0$. Applying the Laplace transformation to (1) we obtain a system of ordinary linear equations in y_i's, which gives

$$y_n(s) = k^n/[E_{n+1} - (k + g)E_n + kgE_{n-1}], \tag{2}$$

E_n being a continuant, i.e., a determinant $\|a_{i,j}\|_n$ with

$$a_{i,i} = s + k + g,\ a_{i,i+1} = g,\ a_{i-1,i} = k,\ a_{i,j} = 0 \text{ for } |j - i| > 1. \tag{3}$$

Expanding E_{n+1} according to the elements of the last column we have

$$E_{n+1} = (s + k + g)E_n - kgE_{n-1}. \tag{4}$$

By (2) and (4) we obtain

$$y_n(s) = k^n/(sE_n). \tag{5}$$

Solving (4) as a finite difference equation in E_n we have[4]

$$E_n = [(kg)^n/(\sigma^2 - 1)]^{1/2}[(\sigma + \sqrt{\sigma^2 - 1})^{n+1} - (\sigma - \sqrt{\sigma^2 - 1})^{n+1}]/2,$$

where

$$\sigma = (s + k + g)/[2(kg)^{1/2}] = \cos\theta. \tag{6}$$

But $\operatorname{arccosh} \sigma = i\theta = \log_e (\sigma + \sqrt{\sigma^2 - 1}) = -\log_e (\sigma - \sqrt{\sigma^2 - 1})$. Therefore,

$$E_n = (kg)^{n/2} \sin (n\theta + \theta)/\sin \theta. \tag{7}$$

The factor in θ is here a standard form of the Tchebychef polynomial of 2nd kind.[5] If $|\sigma| > 1$ we take $\sigma = \cosh \theta$ and replace in (7) sin by sinh. By (7), $E_n = 0$ for

$$\sigma = \cos (i\pi\nu) \text{ where } \nu = 1/(n + 1), i = 1, 2, \ldots, n. \tag{8}$$

Therefore, E_n considered as a polynomial in s has its n roots[6] $s = -\bar{k}_i$, where

$$\bar{k}_i = k + g - 2(kg)^{1/2} \cos (i\pi\nu) > 0. \tag{9}$$

By (3) the highest term of E_n in s is s^n. Therefore

$$E_n = \Pi_{i=1}^{i=n} (s + \bar{k}_i). \tag{10}$$

Consequently by (5) and by known theorems of Laplace transformation:[1, 2]

$$Y_n(t)/k^n = 1^* \exp (-\bar{k}_1 t)^* \exp (-\bar{k}_2 t)^* \ldots ^* \exp (-\bar{k}_n t) = \tag{11}$$

$$(\Pi_{i=1}^{i=n} \bar{k}_i)^{-1} - \sum_{i=1}^{i=n} [\bar{k}_i \Pi_{j=1, j\neq i}^{j=n} (\bar{k}_j - \bar{k}_i)^{-1} \exp (-\bar{k}_i t), \tag{12}$$

$$Y_n(t) = P - (2\sqrt{g/k})^{1-n} e^{-(k+g)t} \sum_{i=1}^{i=n} H_i e^{\tau_i}/c_i, \tag{13}$$

where * stays for convolution, $1^* \ldots = \int_0^t \ldots dt$ and

$$c_i = 1 + (g/k) - 2(g/k)^{1/2} \cos (i\pi\nu),\ 1/P = \Pi_{i=1}^{i=n} c_i, \tag{14}$$

$$1/H_i = \Pi_{j=1, j\neq i}^{j=n} [\cos (i\pi\nu) - \cos (j\pi\nu)],\ \tau_i = 2t\sqrt{kg} \cos (i\pi\nu). \tag{15}$$

From (14) we have by a known formula[7]

$$P = [1 - (g/k)]/[1 - (g/k)^{n+1}]. \tag{14'}$$

From (13) we see that $Y_n(t) \to P$ when $t \to \infty$, because $(k + g)t - \tau_i = \bar{k}_i t > 0$. A simple formula for H_i is obtained as follows: from (7) (9) (10) and (6) we have

$$\Pi_{j=1}^{j=n} [\cos \theta - \cos (j\pi\nu)] = 2^{-n} \sin (n\theta + \theta)/\sin \theta, \tag{16}$$

which could be derived also from known relations for trigonometric products.[7] Putting $\theta = i\pi\nu + \delta$ in (16) we see from (15) that H_i is the limit for $\delta = 0$ of

$$[\cos (i\pi\nu + \delta) - \cos (i\pi\nu)]/\Pi_{j=1}^{j=n} [\cos (i\pi\nu + \delta) - \cos (j\pi\nu)] =$$
$$2^n[\ldots] \sin (i\pi\nu + \delta)/\sin [(n + 1)(i\pi\nu + \delta)],$$

where [...] is in both numeratots the same. Consequently

$$H_i = -(-1)^i 2^n \nu \sin^2 (i\pi\nu).$$

Second Process.—Let $k_i = k (1 \leq i \leq n)$, $g_i = g$ $(0 \leq i \leq n - 2)$, $k_{n+1} = g_{n-1} = 0$. We have now

$$y_n(s) = k^n/(sG_n) \text{ where } G_n = E_n - gE_{n-1}. \tag{17}$$

$Y_n(t)$ is still given by (11) (12) if the roots of G_n are called $s = -\bar{k}_i$ $(i = 1, \ldots, n)$. Since[2] $(E_n)_{s=0} = k^n/P$ by (10), (9), (14), from (14′) we have $(G_n)_{s=0} = k^n$. Therefore:

$$\bar{k}_1\bar{k}_2 \ldots \bar{k}_n = k^n, \tag{18}$$

which reduces by one the number of roots to be calculated. We write:

$$\bar{k}_i = k + g - 2(kg)^{1/2} \cos \theta_i, \; (i = 1, \ldots, n), \tag{19}$$

where the θ_i's are n values of θ[cf. (6)]. We locate these values so that the calculation of the roots may be carried out by routine methods. From (7) (17) we have

$$G_n = -(-1)^i(kg)^{n/2} \text{ for } \theta = (i - 1)\pi/n, \; (i \geq 2), \tag{20}$$

$$G_n = (-1)^i(kg)^{n/2} \, (g/k)^{1/2} \text{ for } \theta = i\pi/(n + 1), \; (i \geq 1), \tag{21}$$

$$G_n = (kg)^{n/2} \, [n + 1 - n(g/k)^{1/2}] \text{ for } \theta = 0. \tag{22}$$

Therefore, if $g/k < (1 + n^{-1})^2$, so that $[\ldots]$ in (22) is > 0 we have:

$$(i - 1)\pi/n < \theta_i < i\pi/(n + 1), \qquad (i = 1, \ldots n), \tag{23}$$

which gives a particularly good location of the n roots for large n. If $[\ldots]$ in (22) is < 0, we must take (23) for $2 \leq i \leq n$, which locates only $n - 1$ roots, however, the nth root may be calculated from (18).[8] From (18) and (12) we see that $Y_n(t) \to 1$ as $t \to \infty$.

Third Process.—Let $k_i = k(1 \leq i \leq n + 1)$, $g_i = g(0 \leq i \leq n - 1)$. We have now $y_n(s) = k^n/G_{n+1}$. The $n + 1$ roots of G_{n+1} are located by equations (18) to (23) in which, however, n must be changed into $n + 1$. By known methods[1, 3] we have now:

$$\begin{aligned} Y_n(t) &= k^n \exp(-\bar{k}_1 t) * \exp(-\bar{k}_2 t) * \ldots * \exp(-\bar{k}_{n+1} t) \\ &= k^n \textstyle\sum_{i=1}^{n+1} [\Pi_{j=1, j\neq i}^{n+1} (\bar{k}_j - \bar{k}_i)]^{-1} \exp(-\bar{k}_i t). \end{aligned}$$

Since $\bar{k}_i > 0$, we have $Y_n(t) \to 0$ as $t \to \infty$.

More General Processes.—If in the previous processes some g_i's with $i \leq n - 2$ are zero, it is easy to see[2] that the corresponding y_n is a product of some y_n's of the previous three types, therefore, the $Y_n(t)$ is a convolution of some Y_n's of the same types.

[1] For bibliography on the subject cf., e.g., Opatowski, I., these PROCEEDINGS, **28**, 84–88 (1942). The probabilities $k_i dt + o(dt)$ and $g_i dt + o(dt)$ are conditional; they give the probabilities of transitions within the interval of time $(t, t + dt)$ under the assumption that at the time t the states $i - 1$ and $i + 1$ existed, respectively.

[2] Opatowski, I., *Bull. Math. Biophysics*, **7**, 161–180 (1945); **8** (1946).
[3] Bateman, H., *Proc. Phil. Soc. Cambridge*, **15**, 423–427 (1910).
[4] Scott, R. F., *Messenger of Math.*, **8**, 131–138 (1879).
[5] Szegö, G., *Orthogonal Polynomials*, New York, 1939, pp. 3, 118.
[6] Cf. Muir, T., *Educational Times*, **42**, 95–96 (1885).
[7] Cauchy, A. L., *Algebraische Analyse*, Berlin, 1885, pp. 381–382.
[8] Instead of (18) the relation $\Sigma_{i=1}^{i=n} k_i = nk + (n-1)g$ can also be used to calculate one of the roots when all the remaining roots are known.

GENERALIZED ARC-SETS

By A. D. Wallace

Department of Mathematics, The University of Pennsylvania

Communicated October 15, 1945

The purpose of this note is to describe a type of strengthened extension set. These sets may be regarded as generalizations of the A-sets introduced by W. L. Ayres and G. T. Whyburn and as homotopic analogs of sets studied by the latter in his work on cyclic elements of higher order.

In what follows H will denote a compact (= bi-compact) Hausdorff space and *subspace* will mean *closed subset of H*. As usual S_n is the unit sphere of $(n + 1)$-dimensional euclidean space. A subspace is a C_n if all mappings of it into S_n are inessential (and hence null-homotopic). By a T_n is meant a subspace all of whose closed subsets are of type C_n. Thus, in the sense of homotopy, a T_n is trivial in dimension n, at least in so far as spheres are concerned. Since an AR can be contracted over itself to a point it must be a C_n for all n. A p-cell is a T_n for $p \leq n$. But S_n is not a C_n and hence not a T_n.

If M is a subspace then M is termed an *extension set in dimension n* (or a J_n) provided that for each subspace X and mapping $f: M \cdot X \to S_n$ there is an f; $X \to S_n$ such that $f = f|M \cdot X$. A subset of H is a B_n if it is not cut by any T_n and is maximal relative to this property. This last concept is due to G. T. Whyburn. The sets T_n are homotopic analogs of sets used by Whyburn to define cyclic elements of higher order in the same manner in which the sets B_n are defined. It is to be emphasized that extension sets are defined *relative* to H and the same remark applies to the sets B_n.

The sets J_n suffer from the serious defect that they may lie wholly within a set of type B_n. For example, let N be defined by the equations $(t \geq 0)$ $x = e^{-t} + |\cos t|$, $y = \sin t$, let L denote the segment joining $(2, 0)$ to $(2, -3)$ and $(2, -3)$ to $(0, -3)$. Finally, if N' is the image of N in the Y-axis and C the circle $x^2 + y^2 = 1$, let H be the union of N, N', L, L' and

C. Then H is a B_0 and C is a J_0. This latter fact follows readily from Theorem 1 below.

Suppose that X is a subspace and f a mapping of X into S_n. Then a closed subset Y of H is an *essential membrane* for f if f can be extended to a mapping of any closed proper subset of Y into S_n but not to a mapping of Y into S_n. This concept is due to W. Hurewicz.

THEOREM I. *In order that a subspace M be a J_n it is n.a.s. that for any closed set X in M and mapping $f: X \to S_n$, each essential membrane for f be contained in M.*

To avoid the difficulty indicated in the example we define a set N to be an A_n provided that

(i) N is a J_n,

(ii) If X is a B_n and $X \cdot N$ is not a T_n then X is a subset of N.

THEOREM II. *The intersection of any family of A_n's is an A_n. Each B_n is an A_n.*

This result is well known if $n = 0$ and H is Peanian. Analogous results hold for sets involving the higher order cyclic elements. In the proof use is made of the fact that J_n's have the same intersection property and that each B_n is a J_n.

THEOREM III. *Let X be a subspace such that H-X is the union of a family of pairwise disjoint open sets the boundary of each of which is a T_n. Then X is an A_n.*

Again when $n = 0$ and H is Peanian this theorem and its converse are well-known results of the cyclic element theory. As to the converse we give an example. Let X be the set $x^2 + y^2 + z^2 = 2$ and U that part of the set $(x^2 + y^2 = 1, 0 \leq z \leq 3)$ which lies *outside* of X. Then, if $H = X + U, H - X = U$ cannot be expressed as the union of a collection of pairwise disjoint open sets (other than the family consisting of U itself) and $\bar{U} - U$ is an S_1 and hence certainly not a T_1. It is easy to see that X is a B_1 and hence a J_1. Here H is an ANR and so locally connected in all dimensions in the strongest possible manner.

The space H is said to be of type V provided that for any pair of closed sets A and B there is a decomposition of H into closed sets A' and B' such that $A \subset A', B \subset B'$, and $A' \cdot B' \cdot (A + B) = A \cdot B$. If H is metric it is of type V but there are spaces of type V which are not metric. The exact position of this axiom in the hierarchy of separation postulates is not known. It is clearly stronger than normality.

THEOREM IV. *If H is of type V then each A_n is an A_{n+1}.*

The proof of this proposition makes use of fact that each T_n is a T_{n+1} and each J_n is a J_{n+1}.

It is clear that H is an A_n and that any T_m is an A_n if $m \leq n$. Also it may be shown that if X is a C_n and M an A_n then $M \cdot X$ is a C_n. It follows that if H is a C_n so also is each A_n and hence each B_n

To generalize the notion of an end-point we say that a subspace E is a T_n-end-element if each neighborhood of E contains a neighborhood whose boundary is a T_n. In this way a point is an end-point if and only if it is a T_0-end-point.

THEOREM V. *Each T_n-end-element is an A_n.*

By a prime $J_n(A_n)$ is meant a $J_n(A_n)$ no closed subset of which is a $J_n(A_n)$ without at the same time being a T_n.

LEMMA 1. *Let A and B be subspaces such that A, B and $A \cdot B$ are T_n's and $A + B$ is a C_n. Then $A + B$ is a T_n.*

This result is no longer valid if the assumption that $A + B$ be a C_n is deleted. In general it can only be said that the union of two T_n's is a T_{n+1}.

LEMMA 2. *If X is an A_n, Z a T_n and $X - Z = U + V$, $U|V$, then the sets $U + Z$ and $V + Z$ are A_n's.*

With the aid of these results it may be shown that

THEOREM VI. *If H is a C_n each prime A_n is a B_n and conversely.*

THEOREM VII. *Let X and Y be subspaces such that $X + Y$ and $X \cdot Y$ are A_n's. Then each of X and Y is an A_n.*

For a Peanian H the original definition of an A_0 was as follows: X is an A-set if it contains each simple continuous arc whose end-points it contains. If now Q_{n+1} is an $(n + 1)$-cell it is clear from Theorem I that an A_n has the property that it contains each Q_{n+1} whose homology boundary it contains. Thus, if in the second example, we add to H the set $(x^2 + y^2 = 1, z = 3)$ then X is no longer an A_1 since it contains the boundary of a Q_2 without containing the Q_2.

LEMMA 3. *If H is a T_n then $H \times (01)$ is a T_{n+1}.*

With the aid of this result and certain others which follow from it we can show that

THEOREM VIII. *If H is a C_n and not a T_n then H contains non-degenerate sets B_{n-1}.*

It should be possible to replace the hypothesis that H is a C_n by the assumption that H is C_{n-1}.

If the mapping f of X into S_n is essential but not essential on any closed proper subset of X then f is said to be *irreducibly essential* on X. It may be shown, for example, that if H admits an essential mapping onto S_n then this mapping is irreducibly essential on some subspace. Parenthetically we interpolate the following: Suppose that H admits an irreducibly essential mapping onto an n-sphere. One might be inclined to conjecture that there exists a subspace with the property that every one of its essential mappings into S_n is irreducibly essential. This need not be true even if H is a finite dimensional metric space (of course compact).

PROPOSITION. *If H admits an irreducibly essential mapping onto S_n then H is not cut by a T_{n-1}.*

As of the present writing the validity of this assertion can only be affirmed in certain highly restricted situations.[1]

[1] A detailed discussion of the sets B_n, C_n, J_n and T_n will appear in the *Trans. Am. Math. Soc.* where pertinent references to the work of Borsuk, Eilenberg, Hurewicz and Whyburn will be given.

TWO MEAN THEOREMS IN HILBERT SPACE

BY KY FAN

THE INSTITUTE FOR ADVANCED STUDY

Communicated November 6, 1945

The ergodic theory and the researches concerning the law of large numbers in the Theory of Probability have demonstrated the importance of the study of the asymptotic behavior of the arithmetical mean

$$\frac{x_1 + x_2 + \ldots + x_n}{n}$$

of a sequence $\{x_n\}$ of points in Hilbert space. In this note we shall give two mean theorems, each of which can be regarded as generalizing the following well-known theorem due to J. von Neumann:[1] For any unitary operator U in the Hilbert space $\mathfrak{H}$ and for any given point x in $\mathfrak{H}$, the arithmetical mean of the n first iterated images of x

$$\frac{Ux + U^2x + \ldots + U^nx}{n}$$

converges strongly to a limit point.

THEOREM I. *Let $\{x_n\}$ be a sequence of points in Hilbert space (real or complex). If there are two constants c_1, c_2 such that the inequalities*

$$||x_{n+1} + x_{n+2} + \ldots + x_{n+m}||^2 - ||x_1 + x_2 + \ldots + x_m||^2 < c_1 m \quad (1)$$

and

$$||x_1 + x_2 + \ldots + x_{n+2}||^2 - 2||x_1 + x_2 + \ldots + x_{n+1}||^2 + ||x_1 + x_2 + \ldots + x_n||^2 > c_2 \quad (2)$$

hold for all positive integers m, n, then the limit

$$\lim_{n \to \infty} \frac{x_1 + x_2 + \ldots + x_n}{n} \quad (3)$$

exists in the sense of strong convergence.

Proof.—We denote as usual by (x, y) the scalar product of two points x, y in $\mathfrak{H}$ and use the following notations:

$$s_n = x_1 + x_2 + \ldots + x_n,$$

$$((x; y)) = (x, y) + (y, x),$$

$$\alpha_k = \sum_{i=1}^{k} ((x_i, x_{k+1})),$$

$$\beta_n = \sum_{k=1}^{n-1} \alpha_k.$$

Obviously the numbers $((x, y))$, α_k, β_n are always real.

It is easy to see that (1) is equivalent to the inequality

$$\left\|\frac{s_n}{n} - \frac{s_{n+m}}{n+m}\right\|^2 < \frac{m}{n(n+m)}\left(\frac{\|s_n\|^2}{n} + \frac{\|s_m\|^2}{m} - \frac{\|s_{n+m}\|^2}{n+m} + c_1\right),$$

which can be written

$$\left\|\frac{s_n}{n} - \frac{s_{n+m}}{n+m}\right\|^2 < \frac{m}{n(n+m)}\left(\frac{1}{n}\sum_{i=1}^{n} \|x_i\|^2 + \frac{1}{m}\sum_{i=1}^{m} \|x_i\|^2 - \frac{1}{n+m}\sum_{i=1}^{n+m} \|x_i\|^2 + c_1\right) + \frac{m}{n(n+m)}\left(\frac{\beta_n}{n} + \frac{\beta_m}{m} - \frac{\beta_{n+m}}{n+m}\right). \quad (4)$$

But the hypothesis (1) implies also the boundedness of $\|x_i\|^2$. Let c_3 be a constant such that

$$\|x_i\|^2 < c_3 \qquad (i = 1, 2, 3, \ldots). \quad (5)$$

Then the expression in the first bracket on the right-hand side of (4) is less than $c_4 = 2c_3 + c_1$. Hence, we have

$$\left\|\frac{s_n}{n} - \frac{s_{n+m}}{n+m}\right\|^2 < \frac{c_4}{n} + \frac{m}{n(n+m)}\left(\frac{\beta_n}{n} + \frac{\beta_m}{m} - \frac{\beta_{n+m}}{n+m}\right)$$

or

$$\left\|\frac{s_n}{n} - \frac{s_{n+m}}{n+m}\right\|^2 < \frac{c_4}{n} + \frac{m}{n+m}\frac{\beta_n}{n^2} + \left(\frac{m}{n+m}\right)^2 \frac{\beta_m}{m^2} - \frac{m}{(n+m)^2}\frac{\beta_{n+m} - \beta_m}{n}. \quad (6)$$

Now, in virtue of the Schwarz inequality and (5), the numbers α_k/k are bounded. If we set

$$\lambda = \limsup_{k \to \infty} \frac{\alpha_k}{k}, \quad (7)$$

we have

$$\frac{\beta_n}{n^2} = \frac{1}{n^2}\sum_{k=1}^{n-1} k\cdot\frac{\alpha_k}{k} < \frac{\lambda}{2} + \epsilon(n), \tag{8}$$

$\epsilon(n)$ being a function of n suitably chosen with $\lim_{n\to\infty} \epsilon(n) = 0$.

Making use of the hypothesis (2) and (5), we can find a positive constant c_5 such that

$$\alpha_{k+1} > \alpha_k - c_5 \qquad (k = 1, 2, 3, \ldots). \tag{9}$$

Consequently we have

$$\beta_{n+m} - \beta_m = \sum_{k=m}^{m+n-1} \alpha_k > n\alpha_m - \frac{n(n-1)}{2} c_5 > n(\alpha_m - nc_5). \tag{10}$$

From (6), (8) and (10), it follows that

$$\left\|\frac{s_n}{n} - \frac{s_{n+m}}{n+m}\right\|^2 < \frac{c_4}{n} + \frac{m}{n+m}\left(\frac{\lambda}{2} + \epsilon(n)\right) + \left(\frac{m}{n+m}\right)^2\left(\frac{\lambda}{2} + \epsilon(m)\right) - \frac{m}{(n+m)^2}(\alpha_m - nc_5). \tag{11}$$

Similarly, if l is a positive integer not greater than n, we may write the inequality[2]

$$\left\|\frac{s_l}{l} - \frac{s_{n+m}}{n+m}\right\|^2 < \frac{c_4}{l} + \frac{n+m-l}{n+m}\left(\frac{\lambda}{2} + \epsilon(l)\right) + \left(\frac{n+m-l}{n+m}\right)^2\left(\frac{\lambda}{2} + \epsilon(n+m-l)\right) - \frac{n+m-l}{(n+m)^2}(\alpha_m - nc_5). \tag{12}$$

According to (11), (12) and

$$\left\|\frac{s_n}{n} - \frac{s_l}{l}\right\|^2 \leqq 2\left\|\frac{s_n}{n} - \frac{s_{n+m}}{n+m}\right\|^2 + 2\left\|\frac{s_l}{l} - \frac{s_{n+m}}{n+m}\right\|^2, \tag{13}$$

we obtain for $l \leqq n$:

$$\left\|\frac{s_n}{n} - \frac{s_l}{l}\right\|^2 < \frac{2c_4}{n} + \frac{2m}{n+m}\left(\frac{\lambda}{2} + \epsilon(n)\right) + 2\left(\frac{m}{n+m}\right)^2\left(\frac{\lambda}{2} + \epsilon(m)\right) - \frac{2m}{(n+m)^2}(\alpha_m - nc_5) + \frac{2c_4}{l} + \frac{2(n+m-l)}{n+m}\left(\frac{\lambda}{2} + \epsilon(l)\right) + 2\left(\frac{n+m-l}{n+m}\right)^2\left(\frac{\lambda}{2} + \epsilon(n+m-l)\right) - \frac{2(n+m-l)}{(n+m)^2}(\alpha_m - nc_5). \tag{14}$$

If now we keep n, l fixed and let $m \to \infty$ through a suitable sequence of integral values such that $(\alpha_m/m) \to \lambda$, (14) becomes

$$\left\|\frac{s_n}{n} - \frac{s_l}{l}\right\|^2 \leqq 2c_4\left(\frac{1}{n} + \frac{1}{l}\right) + 2\epsilon(n) + 2\epsilon(l). \tag{15}$$

As this inequality is symmetric with respect to n, l, it remains true even if we cease to assume $l \leqq n$. This inequality proves the existence of the strong limit (3).

THEOREM II. *Let $\{x_n\}$ be a sequence of points in Hilbert space (real or complex). If there is a constant c_1 such that the inequality (1) is fulfilled for all positive integers m, n and if the limit*

$$\lim_{n\to\infty} \frac{||x_1 + x_2 + \ldots + x_{n+1}||^2 - ||x_1 + x_2 + \ldots + x_n||^2}{n} \tag{16}$$

exists, then the strong limit (3) exists also.

Proof.—Using the same notations as above, we have still the inequality (6), which is, in fact, a consequence of the hypothesis (1) alone. But, according to the present second hypothesis, we have this time, instead of (7) and (8):

$$\lambda = \lim_{k\to\infty} \frac{\alpha_k}{k} \tag{17}$$

and

$$\frac{\beta_n}{n^2} = \frac{\lambda}{2} + \epsilon(n) \tag{18}$$

with $\lim_{n\to\infty} \epsilon(n) = 0$.

By setting

$$\gamma(n, m) = \frac{m}{(n+m)^2}\,\frac{\beta_{n+m} - \beta_m}{n}, \tag{19}$$

we infer from (17) that

$$\lim_{m\to\infty} \gamma(n, m) = \lambda \tag{20}$$

for any fixed value of n.

According to (18), (19), the inequality (6) may be written

$$\left\|\frac{s_n}{n} - \frac{s_{n+m}}{n+m}\right\|^2 < \frac{c_4}{n} + \frac{m}{n+m}\left(\frac{\lambda}{2} + \epsilon(n)\right) + \left(\frac{m}{n+m}\right)^2\left(\frac{\lambda}{2} + \epsilon(m)\right) - \gamma(n, m). \tag{21}$$

Similarly, for any positive integer $l \leqq m$, we have

$$\left\|\frac{s_l}{l} - \frac{s_{n+m}}{n+m}\right\|^2 < \frac{c_4}{l} + \frac{n+m-l}{n+m}\left(\frac{\lambda}{2} + \epsilon(l)\right) + \left(\frac{n+m-l}{n+m}\right)^2\left(\frac{\lambda}{2} + \epsilon(n+m-l)\right) - \gamma(l, n+m-l). \quad (22)$$

Now, using the inequality (13), we can deduce from (21), (22) an inequality which is like (14) and which can be again reduced to (15) by passing to the limit $m \to \infty$. Our theorem II is thus proved.

We terminate this note with the following remark from the point of view of the Theory of Probability. If we consider a sequence of chance variables instead of a sequence of points in Hilbert space, a similar argument will furnish two theorems giving new sufficient conditions for the validity of the law of large numbers. The results thus obtained contain as a particular case an important theorem due to A. Khintchine,[3] which asserts that any stationary sequence of chance variables obeys the law of large numbers.

[1] Neumann, J. von, these PROCEEDINGS, 18, 70–82 (1932). See also Hopf, E., *Ergodentheorie*, Berlin, 1937, p. 23.

[2] We use here the inequality $\alpha_{n+m-l} - lc_4 \geqq \alpha_m - nc_4$, which follows readily from (9) and $l \leqq n$.

[3] Khintchine, A., *Rec. Math. Moscow*, **40**, 124–128 (1933).

INDEX